ACQUIRING KNOWLEDGE FOR AFFECTIVE STATE RECOGNITION IN SOCIAL MEDIA

by

Ashequl Qadir

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

School of Computing

The University of Utah

August 2016

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Ashequl Qadir**

has been approved by the following supervisory committee members:

Ellen Riloff	, Chair	**05/13/2016** Date Approved
Feifei Li	, Member	**05/13/2016** Date Approved
Jeffrey Phillips	, Member	**05/13/2016** Date Approved
Vivek Srikumar	, Member	**05/13/2016** Date Approved
Marilyn Walker	, Member	**05/13/2016** Date Approved

and by **Ross Whitaker**, Chair/Dean of

the Department/College/School of **Computing**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Over the last decade, social media has emerged as a revolutionary platform for informal communication and social interactions among people. Publicly expressing thoughts, opinions, and feelings is one of the key characteristics of social media. In this dissertation, I present research on automatically acquiring knowledge from social media that can be used to recognize people's *affective state* (i.e., what someone feels at a given time) in text. This research addresses two types of affective knowledge: 1) hashtag indicators of emotion consisting of emotion hashtags and emotion hashtag patterns, and 2) affective understanding of similes (a form of figurative comparison).

My research introduces a bootstrapped learning algorithm for learning hashtag indicators of emotions from tweets with respect to five emotion categories: Affection, Anger/Rage, Fear/Anxiety, Joy, and Sadness/Disappointment. With a few seed emotion hashtags per emotion category, the bootstrapping algorithm iteratively learns new hashtags and more generalized hashtag patterns by analyzing emotion in tweets that contain these indicators. Emotion phrases are also harvested from the learned indicators to train additional classifiers that use the surrounding word context of the phrases as features. This is the first work to learn hashtag indicators of emotions.

My research also presents a supervised classification method for classifying affective polarity of similes in Twitter. Using lexical, semantic, and sentiment properties of different simile components as features, supervised classifiers are trained to classify a simile into a positive or negative affective polarity class. The property of comparison is also fundamental to the affective understanding of similes. My research introduces a novel framework for inferring implicit properties that 1) uses syntactic constructions, statistical association, dictionary definitions and word embedding vector similarity to generate and rank candidate properties, 2) re-ranks the top properties using influence from multiple simile components, and 3) aggregates the ranks of each property from different methods to create a final ranked list of properties. The inferred properties are used to derive additional features for the supervised classifiers to further improve affective polarity recognition. Experimental results show substantial improvements in affective understanding of similes over the use of existing sentiment resources.

To my loving wife, Snigdha

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I am grateful to my advisor, Professor Ellen Riloff, without whom I would not be the researcher I am today. Ellen helped me realize my potential as a researcher, inspired me to be creative in research problem solving, and taught me how to always keep the big picture in perspective. Whenever I felt I was stuck, after meeting with Ellen, I always came out of her office with my mind full of new ideas and excitedly ready to explore dimensions I did not think of before. Throughout my PhD years, she guided me to practice and uphold the highest standards, principles, and quality in research. I cannot think of an advisor who could have possibly been more inspiring, motivating and encouraging.

I am thankful to Professor Marilyn Walker for being actively involved in the major parts of this research, and for her many ideas and suggestions that have elevated the quality and added analytic rigor to this work. I thank Professor Vivek Srikumar for his insightful suggestions, comments, and for brainstorming with me over occasional meetings. I would also like to thank my other committee members, Professor Jeff Phillips and Professor Feifei Li, for their helpful inputs during the proposal defense to improve this work.

I have been tremendously lucky to have three amazing mentors and colleagues, Dr. Patrick Pantel, Dr. Michael Gamon, and Dr. Pablo Mendes, who I have had the pleasure to work with and learn from during my internships. At Microsoft Research, I was amazed by Patrick's unique ability to inspire, and Michael's never-depleting patience in mentoring and guiding me. To this date, I greatly miss brainstorming research ideas with Pablo during my IBM research internship at Almaden. I must also thank my master's degree advisor, Dr. Constantin Orasan, for introducing me to the research world of Natural Language Processing.

Where I stand today would not have been possible without the love and support of my wife, Fatema Binte Ahad (Snigdha). She believed in me, and made me believe how far I can go and how much I can achieve. She helped me stay in perspective, and always kept inspiring me with her own academic accomplishments, to go beyond the limits. I am grateful to my parents, Md. Abdul Kader Miah and Asma Kader, for encouraging me throughout my life and for being supportive of whichever academic and career path I chose. I am grateful to my parents-in-law, Md. Nur Ahad and Afia Ahad, for their countless prayers

for my success. I am thankful to my sisters, Shormin Momtaj and Shamima Momtaz, and my brothers-in-law, Sakhawat Abul Kalam Basir and Golam F. Mainuddin, for their unconditional love and support.

Last but not the least, I have had the pleasure to spend time with a good number of people during my PhD years. They have been my friends, colleagues, and patient listeners of my research ideas. I will remember my numerous chats with Ruihong Huang, Lalindra De Silva, Haibo Ding, Nathan Gilbert, Nic Bertagnolli, Xingyuan Pan, and Youngjun Kim at the NLP Lab. I would also like to thank Tobin Yehle, Tao Li, Jie Cao, Annie Cherkaev, and Neetu Pathak for their valuable feedback and comments at my research presentations.

Finally, I thankfully acknowledge the funding agencies without whose financial support this research would not have been possible. This research work was supported in part by the National Science Foundation (NSF) under grants IIS-1450527 and IIS-1302668, and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/ NBC) contract number D12PC00285. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBE, NSF, or the U.S. Government.

CHAPTER 1

INTRODUCTION

With the advancement of Artificial Intelligence and Natural Language Processing, intelligent computer systems are gradually getting better at understanding language and acquiring knowledge from written text. Despite this, the capabilities of modern computers in understanding people's feelings in different situations are still limited. What an individual feels at a given time is referred to as an affective state for that individual (Albrecht et al., 2010). Automatically recognizing affective states requires explicit knowledge of the emotional cues that people exhibit in situations, as well as knowledge to understand various properties of situations that evoke people's feelings. This dissertation addresses how affective knowledge can be automatically acquired from social media text.

One of the advantages people have over computers is that people constantly learn from their experience. They observe how other people feel and react in different situations. The affective knowledge required to understand these situations is naturally acquired by people over time. On the other hand, computer systems are artificially built. To endow computers with the ability to understand affective situations in text, they must be provided with the required affective knowledge, and need to be explicitly taught what to look for when information about the situations is made available to them. The types of affective knowledge computers will need depends on how and where the information about the situations is depicted in text. If the situations are described in social media text, then emotion indicators that are specific to social media become one of the prominent types of affective knowledge that the computer systems can use. If the situations are described in figurative expressions, the computer systems will need to know how to interpret figurative language.

Over the last decade, social media has revolutionized how people stay in touch with each other, interact with each other for informal communication purposes, and share their thoughts and feelings publicly. Some of the most popular social media platforms are weblogs, social networking sites, and microblogs. Twitter, a microblogging platform, is particularly well known for its use by people who instantly express thoughts within a limited length of 140 characters. These status updates, known as tweets, are often emotional and frequently

describe a tweet writer's affective state. The use of emotion cues and figurative comparisons are common in Twitter, and they are among the popular ways people express what they feel or describe emotional situations. The knowledge of the polarity or emotion in individual words is beneficial, but often not sufficient for recognizing people's affective states.

Automatically recognizing affective states, and correspondingly affective tweets, is a challenging task partly because of the many different and unique ways people express their affective states in writing. Making progress on affective state recognition may require acquiring many different types of affective knowledge. This research addresses two such types: 1) hashtag indicators of emotions that can be learned from tweets, and 2) affective understanding of similes (a form of figurative comparison).

The hashtag indicators addressed in this research consist of hashtags, more generalized hashtag patterns, and phrases that represent a writer's emotion in tweets. They are learned using a bootstrapped learning algorithm with respect to five emotion categories: Affection, Anger/Rage, Fear/Anxiety, Joy, and Sadness/Disappointment. For affective understanding of similes, this research presents methods to classify positive and negative affective polarity of similes. The property of comparison in a simile is also an important contributing factor to its affective understanding. This research also presents the first computational framework to automatically infer implicit properties in similes, and shows that the inferred properties can improve affective polarity classification of similes.

1.1 Hashtag Indicators of Emotion as Affective Knowledge

In writing, people express their feelings in many ways. A token or expression that indicates the writer's emotional state in text can be referred to as an emotion indicator. In the context of social media, especially Twitter (but not limited to), one example of such indicators is an emotion hashtag (e.g., #feelinghappy, #noonelovesme). Hashtags are common in Twitter. A study by Wang et al. (2011) found that 14.6% of all tweets in their sample contained at least one hashtag.

The use of hashtags is a distinctive characteristic of social media. It is a community-created convention for providing meta-information about a post, first started in Twitter and later adopted in other social media platforms (e.g., Facebook, Tumblr). They are created by adding the '#' symbol as a prefix to a word or a multiword phrase that consists of concatenated words without whitespace (e.g., #hashtagsarefun). People use hashtags in many ways, for example, to represent the topic of a tweet (e.g., #graduation), to convey additional information (e.g., #mybirthdaytoday), or to convey an emotion or affective state (e.g., #pissedoff).

Figure 1.1 shows some examples of real tweets from Twitter. In the first example, the writer expressed AFFECTION for her mom. When it comes to the individual words in the tweet, there are words with positive polarity such as "awesome" and "inspiration". But these words are not directly associated with AFFECTION. The second and third examples express the writers' ANGER and SADNESS, respectively. These two tweets do not have any individual word that can be directly tied to a negative emotion, let alone a specific emotion such as ANGER or SADNESS.

However, all three tweets in Figure 1.1 have hashtags that can be easily associated with specific emotions. For example, a person would normally understand that #loveyoumom indicates AFFECTION, #angryashell indicates ANGER, and #foreveralone indicates SADNESS. To understand the writers' affective states in these tweets, hashtags associated with the corresponding emotions will be valuable affective knowledge for a computer.

There is no set convention for how hashtags should be created. In addition to single word hashtags (e.g., #angry), there are also hashtags that are multiword phrases (e.g., #lovehimsomuch), elongated terms (e.g., #yaaaaay,#goawaaay), creatively spelled hashtags (e.g., #only4you, #YoureDaBest), acronyms (e.g., #lol, #wth), etc. To create a repository of emotion hashtags, one option could be to analyze a large sample of tweets and manually categorize the hashtags in these tweets into emotion categories. But manually analyzing large samples of tweets would require substantial time and effort. Also, people create new hashtags with these stylistic variations everyday. Manually created emotion hashtag repositories with one-time effort will not include novel hashtags. Alternatively, if a lexicon of emotion hashtags is artificially created by trivially adding the '#' symbol as

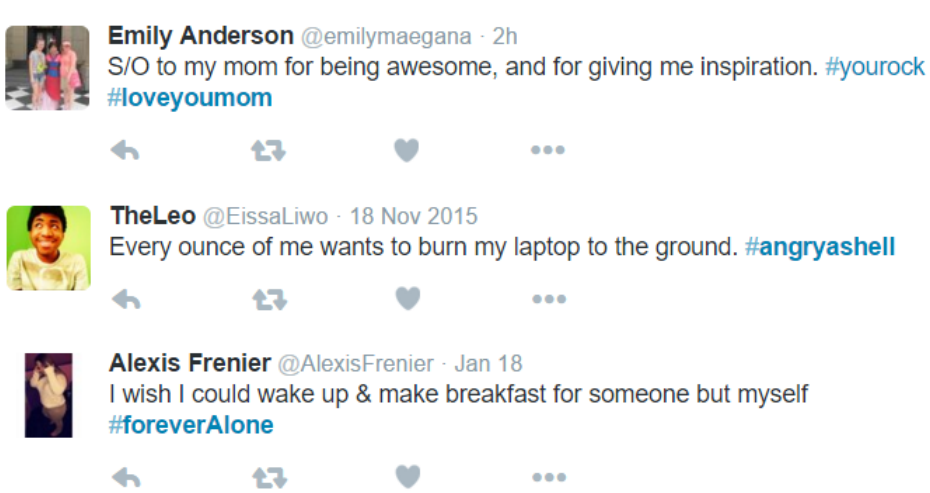

Figure 1.1: Examples of tweets with emotion.

a prefix to the terms in an existing emotion lexicon, similar limitations will still remain. This challenge presents the need for automatic methods that will not require substantial manual effort, can easily be adopted to learn new hashtags not previously seen before, and can continually update the repositories when needed.

A second observation is that different hashtags can have a common prefix word or phrase which is often sufficient to understand the writer's emotion. For example, both #scaredofexamtomorrow and #scaredofghosts have the prefix "scared of", which directly indicates that the writer felt FEAR/ANXIETY. Knowledge of these prefix patterns can allow for generalization to provide additional coverage over specific hashtags. Moreover, if the '#' symbol is stripped off from the emotion hashtags, and the hashtags are expanded into phrases, the resulting expressions become emotion phrases that can be used to understand the writer's emotion in the body of a tweet. The research in this dissertation presents methods to automatically learn these emotion hashtags, emotion hashtag patterns, and emotion phrases as hashtag indicators of emotion with respect to five emotion categories: AFFECTION, ANGER/RAGE, FEAR/ANXIETY, JOY, and SADNESS/DISAPPOINTMENT.

One of the primary applications of the hashtag indicators of emotions is in recognizing people's affective states in tweets. The acquired affective knowledge can also be used for affective text classification in other social media platforms. For example, hashtags have become a common phenomenon in Facebook and Tumblr. Whenever a Facebook or Tumblr post would contain an emotion hashtag, or a hashtag that matches one of the emotion hashtag patterns, the hashtag or the pattern can be used to predict the writer's emotion in the post. The emotion phrases that can be extracted from the hashtags and hashtag patterns are not specific to only social media text, and can also be used for affective state recognition in other text genres such as emails or personal narratives.

For learning the hashtag indicators automatically, I present a bootstrapped learning framework. The learning algorithm requires a small number of seed emotion hashtags per emotion category and a large collection of unlabeled tweets. The initial collection of seed hashtags is small enough (e.g., five hashtags per category) that it can be created with little manual effort. The tweets that contain the seed hashtags supply the learning framework with training instances to train a supervised classifier for emotion classification in tweets. The learning algorithm then decides which hashtags to add to the initial collection by analyzing emotion in tweets where the hashtags appear. The process is then repeated, and the bootstrapping algorithm iteratively grows the initial collection of seed hashtags into a large dictionary of emotion hashtags. The same bootstrapped learning framework is also

extended to learn the more general hashtag patterns. At the end of the learning process, emotion phrases are harvested from the learned hashtags and patterns as the third type of emotion indicator.

The advantage of the bootstrapped learning framework is that it only needs the seed hashtags to jump-start the learning process, and therefore is not limited by the need for substantial manual annotations. The iterative design is also not limited by a fixed set of training instances, which is a typical characteristic of supervised classification. The algorithm can be run on new unannotated data to learn novel hashtags, and at the same time, it can retain any knowledge learned in past iterations, making it well suited for the task. The details of the bootstrapped learning algorithm for learning the hashtag indicators of emotions as affective knowledge are presented in Chapter 3.

1.2 Affective Understanding of Similes

A simile is a figure of speech that explicitly compares two concepts that are different from each other. The explicit comparison can be easily identified from the use of commonly used comparator keywords such as "like" or "as" (Paul, 1970). For example, *"my lawyer is like a shark"* compares two dissimilar entities: "lawyer" and "shark" (Sam and Catrinel, 2006). Similes describe a state or an activity of the subject of the comparison and often contain implicit affective knowledge with respect to how one feels toward these states or activities.

To understand how frequent similes are in Twitter, I analyzed multiple samples of tweets written in English and manually identified the similes in the samples. The findings suggest that a simile can be expected in nearly every 147 random English tweets (0.68%). This percentage increases to nearly 1 simile in every 111 tweets (0.90%) for tweets with positive/negative sentiment. The details of the study will be presented in Chapter 4.

Figure 1.2 shows examples of some real tweets that contain a simile. In the first example, the writer compares holding an iPhone with a bar of soap. A bar of soap is naturally slippery, and the simile indicates that the iPhone has a slippery exterior. With the additional context, it becomes even more clear that this property is problematic for an iPhone because it makes the gripping challenging, highlighting a negative aspect. In the second example, by comparing the mattress with a slab of concrete, the writer means to say that the mattress feels hard. A hard mattress is typically not considered a good mattress, thus the comparison describes a negative quality of the mattress. In the last example, the writer compares his office laptop with a gazelle. Gazelles are known to run fast, and the comparison tells that the laptop is a fast laptop, indicating a positive quality of the laptop.

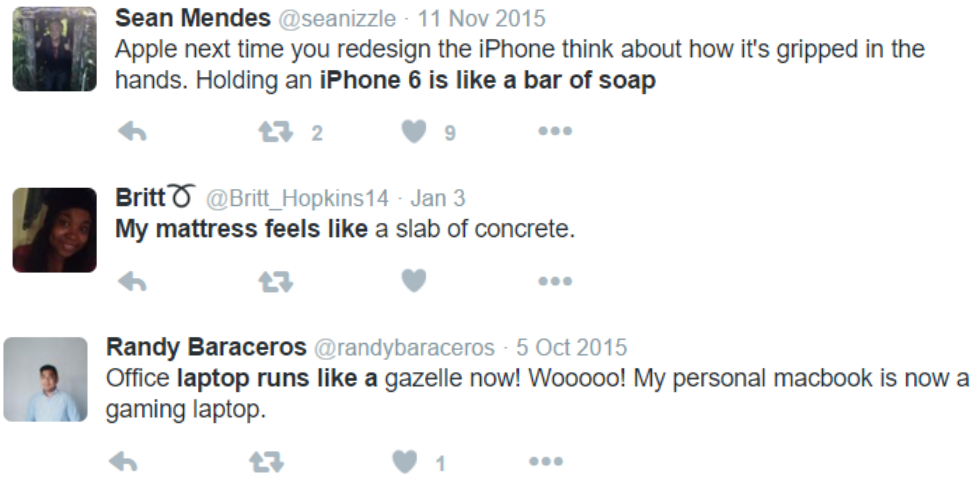

Figure 1.2: Examples of tweets with similes.

Even without any additional context from the surrounding words, people can easily understand whether the affective states evoked by the similes are positive or negative just by looking at the similes alone. But these similes, or even the entire tweets, do not have any easily recognizable positive and negative words, making it challenging for a computer to draw the same conclusion as people effortlessly would. This demonstrates the need for specialized methods for affective understanding of similes.

Similes sometimes explicitly mention the basis of the comparison (e.g., *"John is **brave** like a lion"*), but more often they do not (e.g., *"John was like a lion in battle"*). These two types of similes are commonly known as *closed* and *open similes* (Beardsley, 1981). To estimate the relative frequency of these two types of similes in tweets, I analyzed a random sample of English tweets and found that 92% of the similes in the sample were *open* similes that did not explicitly mention the basis or the common property of comparison. The details of the study will be presented in Chapter 4.

In a *closed* simile, the property is often a direct cue for the affective understanding of the simile. In the example above, "brave" is a positive quality attributed to "John", so the simile has a positive polarity. But since most similes in tweets are *open* similes that do not explicitly mention the basis for comparison, it must be inferred. For example, the simile *"John was like a lion in battle"* contains only neutral words, so being able to infer "bravery" as the implicit property can provide additional information for the affective understanding of the simile. Figure 1.2 demonstrates that it is often easy for people to infer the implicit properties in *open* similes. By comparing with "bar of soap", "slab of concrete", and "gazelle", the writers hint on the properties: *slippery* for the iPhone, *hard*

for the mattress, and *fast* for the laptop, which can be easily understood by a reader.

Affective understanding of similes can be valuable for a computer. Even though a Twitter corpus is used as the source for similes in this research, similes are not exclusive to social media. They are used by people in spoken conversations, arguments, product reviews, and even in commercial jingles (e.g., *"like a good neighbor, State Farm is there"*). The learned affective polarity of the similes can be used to determine someone's subjective perception about the simile's subject in any of these communication mediums.

Inferring the basis of comparison in a simile and being able to recognize the affective polarity is central to natural language understanding. Methods to understand similes can also be valuable for understanding metaphors that have linguistic constructions like predicate nominals (e.g., *"he is a lion"*), and in turn, for affective understanding of these metaphors. Furthermore, associating an inferred property with a comparison subject enables extraction of states and activities of the comparison subject, and the affective polarity of similes can help to understand the affective polarity of these states and activities. For example, *"my room feels like Antarctica"* indicates that my room is cold, and the property "cold" can be associated with the room to describe its state. Knowing the simile has a negative polarity also allows to understand that *"my room feels cold"* describes a negative state of the room.

In this research, I present a supervised classification method for automatically recognizing affective polarity of similes. The supervised classification algorithm uses features that are derived from individual component words of a simile (e.g., subject, object, or verb of the comparison). These features reflect lexical, semantic, and sentiment properties of the component words. The goal of the classifiers is to assign a positive, negative, or neutral label to a simile. As the training instances, the classifiers use manually labeled data that are small in size but of high quality, and also large training data that are automatically labeled but can be expected to contain some noise.

I also present a framework for automatically inferring implicit properties in similes. Using methods that use syntactic structures, measures of statistical association, dictionary definitions of simile component words, and word embedding vector similarity, candidate properties are first generated and ranked. The top properties are then reranked using influence from multiple simile components. Finally, the ranks for each property from different methods are aggregated to produce a final ranked list of properties. The details of the supervised classifiers for affective polarity recognition and the framework for implicit property inference are presented in Chapters 5 and 6.

1.3 Dissertation Claims and Research Contributions

The claims I put forth in this dissertation are:

> *Claim#1: Hashtag indicators of emotions can be automatically learned from tweets using a bootstrapped learning framework.*

I propose a bootstrapping algorithm to automatically learn hashtag indicators for specific emotions using only a small number of seed emotion hashtags and a large collection of unlabeled tweets. The seed emotion hashtags can be acquired with minimal human efforts, and the learning algorithm iteratively builds a repository of emotion hashtags and more general emotion hashtag patterns. The learning framework does not use a fixed set of training instances that are typically used in a supervised classification solution. Instead, it automatically labels new number of training instances in each bootstrapping iteration, allowing it to identify new emotion hashtags and patterns that were unknown in previous iterations. At the end of the bootstrapped learning, emotion phrases are also harvested from the learned emotion hashtags and the patterns. The emotion phrases are used to train additional context-based emotion classifiers using context words of the emotion phrases as features, allowing the classifiers to reliably use the emotion phrases in context.

The learned hashtag indicators of emotions are evaluated in a tweet emotion classification task. The learned hashtags and hashtag patterns are used to recognize emotion in tweets when the tweets contain one of these hashtags or a hashtag that matches a learned emotion hashtag pattern. Emotion classification results on a data set of emotion tweets suggest that good-quality hashtag indicators of emotions could be automatically acquired using the proposed bootstrapped learning framework.

> *Claim#2: Affective interpretation of similes can be automatically achieved by affective polarity classification of similes and by inferring the implicit properties of open similes.*

Existing sentiment resources are insufficient for affective interpretation of similes because similes do not always have words with a positive or negative polarity. I propose a supervised classification method that derives lexical, semantic, and sentiment properties of individual simile components, and uses them as features for automatically classifying affective polarity of similes. One of the challenges of supervised classification is acquiring labeled instances to train a classifier. To this end, I present experiments with both manually labeled training data, and training data that can be automatically labeled using existing sentiment resources.

The majority of similes in Twitter do not explicitly mention a property that is the basis of comparison in a simile. These properties are important contributing factors in the

affective understanding of similes. I propose a framework that first generates candidate properties from multiple simile components using a variety of methods including syntactic constructions, dictionary definitions, statistical association, and word embedding vector similarity. Using influence from complementary simile components, and by combining the individual candidate ranks from different methods in an aggregated ranking, my proposed method is able to automatically infer implicit properties in similes. The inferred properties can be used as additional lexical features to the supervised classifier to improve affective polarity recognition results. My presented method is the first computational model for inferring implicit properties in *open* similes.

Recognizing people's affective state in text using the learned affective knowledge can be beneficial for many other application areas, for example, to help companies understand how people feel about their products, to assist governments in recognizing growing anger or fear associated with an event, or to help media outlets understand the public's emotional state arising from controversial issues or international affairs. The types of affective knowledge addressed in this research are some examples of tractable cases of sentiment or emotion-bearing expressions that are not recognized well by the current state-of-the-art methods, sentiment analysis tools or sentiment/emotion lexicons. Making progress on affective state recognition may require acquisition of many different types of affective knowledge. The research contributions presented in this dissertation address some of these tractable cases, and aim to open future research avenues in similar areas.

1.4 Guide to This Dissertation

The rest of the chapters in this dissertation are organized as follows:

Chapter 2 discusses related work in sentiment/emotion classification, creation of affective knowledge resources, theoretical and computational models for figurative language understanding, and general semantic knowledge resources and methods for creating these resources automatically.

Chapter 3 presents a bootstrapped learning framework that automatically learns emotion hashtags, emotion hashtag patterns, and emotion phrases. The chapter discusses creation of an evaluation data set containing tweets manually labeled with emotion categories. The chapter also presents emotion classification performance using the learned hashtag indicators of emotion, and comparisons against several baseline systems.

Chapter 4 introduces similes as a source of affective knowledge and discusses different aspects of similes that contribute to the understanding of a simile as a figurative comparison, the factors that play a role in the affective polarity of similes, and the role of different

component words for implicit property inference. The chapter discusses creation of a data set containing similes with manual annotations for positive and negative affective polarity. The chapter also discusses creation of a data set of *open* similes that are manually annotated with their implicit properties.

Chapter 5 presents supervised classification of affective polarity in similes using lexical, sentiment, and semantic features derived from individual simile components. The chapter introduces methods for automatically acquiring training instances for supervised classification, and presents affective polarity classification experiments with both manually and automatically labeled instances.

Chapter 6 presents a framework for inferring implicit properties in *open* similes. It discusses coverage of a variety of candidate property generation methods, and presents methods that exploit the influence of multiple simile components to rank the implicit properties with an aggregated ranking. The chapter also presents evaluation results for implicit property inference, and results demonstrating that the inferred properties improve affective polarity recognition.

Chapter 7 concludes the dissertation by summarizing the research contributions in this work, and discussing avenues for future work.

CHAPTER 2

RELATED WORK

Over the last decade, different social media platforms such as weblogs, microblogs, and social networking sites have emerged as popular virtual mediums of communication where people publicly express their thoughts, feelings, and moods. Among these platforms, the microblogging platform Twitter is particularly well known for its popularity. This has put Twitter at the focus of many Natural Language Processing research works.

The most popular research direction in affective tweet recognition aims at determining the overall positive or negative sentiment polarity of a tweet, or emotions of finer dimensions (e.g., joy, sadness). There are other research works that identify sentiment targets, or identify events that evoke sentiment. Different types of knowledge resources have been found to greatly benefit these tasks, and some of the research works have specifically focused on building these knowledge resources. It has also been found that the figurative uses of language (e.g., metaphors, similes) tend to have a strong correlation with sentiment.

These research areas are closely related to the research presented in this dissertation. The following sections briefly discuss related work in these areas.

2.1 Sentiment and Emotion Classification

Sentiment classification work mainly aims at determining positive and negative sentiment polarities, whereas emotion classification work attempts to recognize more fine-grained emotions such as joy, fear, anger, or sadness. Some of these works determine the overall sentiment or emotion of a message or post. The most common approach is supervised classification with a variety of lexical, syntactic, or sentiment lexicon-based features.

Kouloumpis et al. (2011) classified tweets into positive, negative, and neutral classes using the AdaBoost algorithm with features such as unigrams, bigrams, parts-of-speech features, sentiment lexicon features, and microblog-specific features such as all-capitalized words and character repetitions. Mohammad et al. (2013) used a wide variety of features, such as word N-grams, character N-grams, word capitalization, parts-of-speech, number of hashtags, sentiment lexicon features, punctuation, emoticons, negation features, elongated

words, and word clusters representing message content, as features for a SVM model to classify tweets and SMS into positive, negative, and neutral classes. Chikersal et al. (2015) used linguistic rules to selectively use N-gram features (e.g., N-grams after conjunctions but not the N-grams after disjunctions) for a SVM to classify tweets into positive and negative classes.

For classifying text into more fine-grained sentiment or emotion categories, Davidov et al. (2010) classified tweets into sentiment categories where the sentiment labels correspond to 50 sentiment-indicating hashtags and six mood indicating smileys. They used an algorithm similar to k-nearest neighbors but with supervision from training data, and used features such as N-grams (length ranging from two to five), extraction pattern-based features (the patterns contain high-frequency words but match more important low-frequency content words in a message) and punctuation-based features. Roberts et al. (2012) used SVM classifiers to classify tweets into anger, disgust, fear, joy, love, sadness, and surprise categories using features such as unigrams, bigrams, trigrams, punctuations, WordNet synsets, WordNet hypernyms, LDA topics, significant words determined using pointwise mutual information, etc. Wang et al. (2012) used Naive Bayes classification with features such as unigrams, bigrams, parts-of-speech, and sentiment lexicon features to classify tweets into sad, anger, happy, and fear categories. Thomas et al. (2014) used Multinomial Naive Bayes classifier with unigrams, bigrams, trigrams, and emotion lexicon-based features to classify user reported situations into emotion classes: joy, fear, anger, sadness, disgust, shame, and guilt. They used N-grams more selectively using feature selection methods such as Weighted Log Likelihood, Mutual Information, and Normalized Google Distance.

To overcome the challenge of acquiring manually labeled data, some researchers have collected noisy training data using emoticons and hashtags for supervised classification of sentiment (e.g., Go et al., 2009; Pak and Paroubek, 2010; Purver and Battersby, 2012; Suttles and Ide, 2013). Some methods use a hierarchical classification approach by first classifying emotion vs. nonemotion tweets, and then identifying positive vs. negative tweets or tweets with finer dimensions of emotion (e.g., Barbosa and Feng, 2010; Esmin et al., 2012). Some of the works aim at more specific goals and determine people's sentiment for predicting election outcomes (Tumasjan et al., 2010), stock market fluctuations (Bollen et al., 2011), or identify a specific target or topic of sentiment in a tweet (e.g., a movie or company toward which sentiment is directed) (Jiang et al., 2011).

Other text genres studied in sentiment analysis include customer reviews (e.g., product reviews, movie reviews). For sentiment classification, some of these works explored the

impact of higher order n-grams (Cui et al., 2006), negation phrases (Na et al., 2005), sentiment lexicons (Ohana and Tierney, 2009), different levels of intensity for adjectives (Sharma et al., 2015), differences in intentions and perceptions for writer vs. reader (Maks and Vossen, 2013), etc.

Sentiment or emotion classification has also been done on news articles. Some of these works investigated the temporal change of sentiment with a news topic (Fukuhara et al., 2007), sentiment toward entities mentioned in news articles (Godbole et al., 2007), correlation of stock market price direction with sentiments in financial news articles (Schumaker et al., 2012), emotion from the reader's perspective (Lin et al., 2008), emotion in news headlines (Kozareva et al., 2007; Strapparava and Mihalcea, 2008), etc. Emotion and sentiment analysis has also been performed on suicide notes (Pestian et al., 2012; Xu et al., 2012; Desmet and Hoste, 2013) and emails (Mohammad and Yang, 2011).

Another line of work determines sentiment directed toward different aspects of consumer products (e.g., battery life of a camera, or the service, food, or ambience in a restaurant). Instead of social media text, they mainly work with customer reviews written in e-commerce or opinion websites. Common approaches for extracting aspects of consumer products include association mining (Hu and Liu, 2004), different variations of Latent Dirichlet Allocation (LDA) (e.g., Sentence-LDA (Jo and Oh, 2011), Multi-Grain LDA (Titov and McDonald, 2008), Local LDA (Brody and Elhadad, 2010)), use of semantic relations such as parts and properties (Popescu and Etzioni, 2007), or Wikipedia categories (Fahrni and Klenner, 2008). More recent state-of-the-art systems use sequential tagging techniques (Castellucci et al., 2014), entity recognition techniques (Kiritchenko et al., 2014) or dependency relations (Brun et al., 2014).

There are many other research works in the area of sentiment analysis and opinion mining, such as sentiment analysis in movie reviews, weblogs and news articles, determining contextual polarity, sentiment summary generation, classification work at document or sentence level, use of unsupervised and semisupervised methods, etc., and many more. Comprehensive discussion of many of these works can be found in the survey articles by Liu and Zhang (2012), Medhat et al. (2014), and Pang and Lee (2008).

While the majority of research works above are mainly focused on classification of overall sentiment polarity or finer dimensions of emotion categories in messages and documents, this research learns hashtag indicators of emotions for five emotion categories: Affection, Anger/Rage, Fear/Anxiety, Joy, and Sadness/Disappointment, using a bootstrapped leaning framework. This is the first work to learn hashtags for emotions.

The hashtag indicators are evaluated in a tweet emotion classification task with respect to these five emotion categories. This research also classifies similes from Twitter into positive and negative affective polarity. The similes represent states and activities of the comparison subjects, and the polarity of similes helps to understand people's subjective view towards these states and activities. Affective polarity recognition as well as inferring implicit properties of similes addressed in this research aim to improve affective understanding of similes in text.

2.2 Affective Knowledge Resources

Affective knowledge resources typically contain words, phrases, and knowledge that map to positive/negative sentiment classes or emotion categories of finer dimensions such as anger, sadness, joy, etc. Some of the affective knowledge resources contain entities, relations, events, and knowledge about how these concepts are associated with people's affective states. In this section, related work in these areas is discussed.

2.2.1 Positive/Negative Sentiment Word Lexicons

Sentiment lexicons consist of words with positive and negative polarity and occasionally contain words that are neutral. General Inquirer (Stone et al., 1968) is well known as one of the pioneer resources, consisting of words and their positive/negative semantic orientation. Hu and Liu (2004) created a lexicon of positive and negative words compiled over many years, for sentiment analysis purposes. AFINN-111 (Nielsen, 2011) was manually created and enriched with words from Twitter, and internet slang from web dictionaries. LIWC (Pennebaker et al., 2015) is also manually created, and contains words with respect to both positive/negative emotions, as well as emotions such as anxiety, anger, and sadness.

The MPQA Subjectivity Lexicon contains subjectivity clues that were first acquired by automatic bootstrapped learning of extraction patterns associated with subjectivity (Riloff and Wiebe, 2003), which was then later expanded using a dictionary, and thesaurus (Wilson et al., 2005). Mohammad et al. (2009) automatically identified positive and negative seed words using affix patterns, and expanded them using synonyms from a thesaurus. Lu et al. (2011) combines multiple sources of sentiment information such as general-purpose sentiment lexicons, overall sentiment ratings of documents, thesaurus, and linguistic rules to create a context-aware sentiment lexicon where words in the lexicon are conditioned on different aspects of a given domain. (e.g., the word "large" is bad when it describes a laptop battery, but good when it describes a laptop screen). They combined these different sources of information with constraints, under a constraint optimization framework. Mohammad

et al. (2013) automatically built lexicons of unigrams and bigrams from a large collection of tweets using their pointwise mutual information scores with positive and negative sentiment hashtags and emoticons. SenticNet (Cambria et al., 2010) infers concept polarity of commonsense concepts using *Blending* (performing SVD on a blended matrix created from multiple sources of information consisting of commonsense and affective knowledge) and *Spectral Association* (transference of activation to concepts from the key concepts through short paths or many different paths in a common sense knowledge network).

A popular method for automatically creating lexicons is by iterative bootstrapped learning that begins with a small number of seed instances, gradually expanded into bigger lexicons. Riloff et al. (2003) generated a lexicon of subjective nouns with an iterative algorithm that exploits lexico-syntactic patterns. Kanayama and Nasukawa (2006) leverage context coherency of polar words to expand a sentiment lexicon using text in product discussion boards. Volkova et al. (2013) used a high-precision initial sentiment lexicon and expanded it using tweets with the assumption that words in a tweet have the same polar orientation. Jijkoun et al. (2010) used bootstrapping to generate topic-specific sentiment lexicons. The lexicons contained subjectivity clues, sentiment targets, and their syntactic contexts. Banea et al. (2008) constructed sentiment lexicons for languages with scarce resources, using word similarity with the seed terms determined using pointwise mutual information and latent semantic analysis. Qiu et al. (2011) used a double propagation algorithm to learn opinion words and opinion targets (e.g., topics, product features) with the idea that opinion words and targets are often connected by dependency relations, and information can be propagated between the words and targets back and forth.

More closely related to this research is the work by Wang et al. (2011) who classified Twitter hashtags having positive or negative sentiment polarity, by employing several algorithms that exploit sentiment polarity of tweets containing hashtags, hashtag co-occurrence, and the presence of sentiment lexicon words in hashtags. One major difference between their work and the research presented in this dissertation on learning hashtag indicators of emotions is that my research aims to learn hashtags that are associated with finer dimensions of emotion. My research also generalizes beyond specific hashtags by additionally learning emotion hashtag patterns.

2.2.2 Lexicons of Finer Dimensions of Emotions

Beyond positive and negative sentiments, a number of works aimed to build emotion lexicons of finer dimensions. These dimensions typically capture more specific emotions such as *happiness*, *anger*, *fear*, etc. Yang et al. (2007a) built emotion lexicons from weblogs

using emoticons that correspond to affective states such as happy, sad. They used word collocation strengths with emoticons using a variation of the pointwise mutual information calculation. Depeche Mood (Staiano and Guerini, 2014) used a compositional semantics based approach to create an emotion lexicon from emotion labeled news articles. Fraisse and Paroubek (2014) built a multilingual emotion lexicon using hashtags of emotion words (e.g., anger → #anger, fear → #fear) in Twitter. They use hashtags of seed affective words in English, identify emotion hashtags of other languages that co-occur with the seed hashtags in the same tweet, and learn emotion words from these hashtags to create emotion lexicons in multiple languages.

Mohammad and Turney (2013) used crowdsourcing to build an emotion lexicon with unigrams and bigrams from a thesaurus that are frequent in the Google N-gram corpus, and asked Amazon Mechanical Turk workers to provide emotion information for these terms. Mohammad (2011) also used similar crowdsourcing method to associate colors with words, and matched them with word-emotion associations to determine associations between colors and emotions (e.g., red and black colors have frequent association with negative emotions such as disgust, fear, and sadness). Mohammad (2012a) created an emotion lexicon consisting of unigrams and bigrams by computing pointwise mutual information of these N-grams with different emotion hashtags in Twitter. As this lexicon is created from Twitter data, I directly compare their lexicon with the learned hashtag indicators of emotions from this work in a tweet emotion classification task.

2.2.3 Using WordNet for Building Sentiment and Emotion Lexicons

Several works mine affective knowledge from general purpose semantic knowledge bases such as WordNet (Miller, 1995), and build resources with knowledge about affective states. SentiWordNet 3.0 (Baccianella et al., 2010) determines degrees of positivity, negativity, and neutrality of WordNet synsets using a pipeline of semisupervised learning and Random-walk in the WordNet graph. STEP (Adreevskaia and Bergler, 2006) expands positive and negative sentiment-bearing terms using WordNet relations (e.g., hypernyms, antonyms) and organizes them into fuzzy sentiment categories by calculating a Net Overlap Score from sentiment association of the terms in different runs of their algorithm. Kim and Hovy (2004) expands manually selected positive and negative seed terms using WordNet relations for determining sentiment of opinions.

Godbole et al. (2007) expand positive and negative seed terms using WordNet synonyms and antonyms but determine the trustworthiness of the terms in the expanded set by taking into account path depth and path alternation between positive and negative classes. They

learn lexicons with respect to different domains such as general, health, crime, sports, business, politics, and media. SentiFul (Neviarouskaya et al., 2011) expands core positive and negative sentiment words using WordNet synonym, antonym, and hypernym relations, and also expands them using morphological derivations using affixes. Beyond positive and negative words, WordNet-Affect (Strapparava et al., 2004) labels WordNet synsets representing affective concepts with affective domain labels such as *emotion*, *mood*, *trait*, etc., and expands them with WordNet relations.

These lexicons mainly aim to find terms associated with affective polarity or finer dimensions of affective states, but do not contain social media-specific tokens such as hashtags. The lexicon of hashtag indicators of emotions that I present in this research is different in nature than these resources, and can further add to their contributions in affective tweet recognition.

2.2.4 Specialized Affective Knowledge Resources

Researchers have also dedicated efforts to create affective knowledge resources that are different from sentiment or emotion word lexicons. Using graph propagation techniques, Feng et al. (2013) constructs a connotation lexicon with words that have a positive or negative connotation. For example, nouns such as Einstein, Harvard, and verbs such as nurse, volunteer, have positive connotation because these entities and actions are viewed positively by people. On the other hand, nouns such as Enron, Qaeda, and verbs such as scratch, overcharge, have negative connotation because these entities and actions are viewed negatively by people. The difference between these words and traditional sentiment or emotion words is that these words do not explicitly indicate someone's sentiment or emotional states; rather the associated connotation polarity represents people's stereotypical positive or negative views toward the concepts.

AffectNet (Cambria and Hussain, 2012; Cambria et al., 2015) aligns concepts from ConceptNet (Liu and Singh, 2004) with WordNet-Affect (Strapparava et al., 2004) by creating concept vectors, and then comparing their alignment in vector space (e.g., "birthday party" is aligned with *feeling happy* whereas "being laid off" is aligned with *guilt*). Some of these concept vectors represent state descriptions, but they are different from the states represented by similes that I address in this research, as the states in similes are described through comparison.

There has also been research that creates resources for affective events. AESOP (Goyal et al., 2010) acquires patient polarity verbs (e.g., killed, injured, scammed) using patterns designed to identify verbs that co-occur with stereotypical kind and evil agents, and impart

positive or negative affect on their patients. Tokuhisa et al. (2008) created a corpus of emotion-provoking events by mining the web for sentences that contain pre-selected emotion words, and taking subordinate clauses headed by these emotion words. Vu et al. (2014) created a dictionary of emotion-provoking events by using a similar method, and grouped similar events together in the final dictionary. Li et al. (2014) extracted major life events (e.g., university admission, receiving award, etc.) using congratulations or condolences speech acts in Twitter.

Riloff et al. (2013) used a bootstrapped learning algorithm to learn positive sentiment phrases along with negative situations for recognizing sarcastic tweets by contrasting the sentiment phrases and the situations. These negative situations represent states and activities that are stereotypically viewed negatively by people. Choi et al. (2014) and Choi and Wiebe (2014) created a sense-level benefactive/malefactive (also addressed as good for/bad for) events lexicon for events that have positive or negative effects on entities, by exploiting WordNet relations of event verbs. Deng and Wiebe (2015) builds a system that employs probabilistic soft logic to infer explicit and implicit sentiments toward entities and events in text. Ding and Riloff (2016) learns stereotypically positive and negative events from personal blogs using a semisupervised label propagation algorithm.

These knowledge resources contain knowledge about affective events, which relate to the work presented in this dissertation that focuses on affective understanding of similes. As a simile describes a state or activity of the subject of comparison, an activity described in a simile can be considered a type of affective event. For example, *"dad drives like a snail"* describes how dad drives. However, a major difference is that the affective polarity of an activity in a simile typically depends on the entire comparison. In the above example, "driving" does not have a positive or negative polarity by itself. Rather the polarity is evoked because the driving is compared with a snail's movement.

2.3 Similes and Metaphors

Different forms of figurative language such as metaphors and similes are known to have correlation with sentiment or affective states. In this section, some of the theoretical studies as well as computational models of figurative language with respect to similes and metaphors are discussed.[1]

[1]Note that there are other forms of figurative language such as hyperbole, personification, idiom, irony, etc., which are fundamentally different from metaphors and similes, and therefore not discussed as part of the related work.

2.3.1 Theoretical Models and Studies

Similes and other forms of figurative comparisons have been studied in linguistics and psycholinguistics to understand how people process similes, comparisons, and metaphors, and the interplay among different components of these linguistic forms. These research works typically conduct studies with human participants to verify their hypotheses and models. Glucksberg et al. (1997) presented a theoretical property attribution model of metaphor comprehension where the properties of comparison are selected from the object of comparison and are applied to the comparison subject, and possible dimensions of the subject of comparison are imposed as constraints on the selection process. Chiappe and Kennedy (2000) investigated to what extent the number of properties vary between a metaphor and its simile form, and found that metaphors tend to have more properties than similes. The impacts of semantic dimensions of the comparison subject and property salience have been compared by Gagné (2002). They found that high salience aids simile and metaphor comprehension.

Fishelov (2007) experimented with affective connotation and degrees of difficulty associated with understanding a simile when a simile property is conventional or unconventional, or no property is given, and concluded that having a conventional property and having an explicit property are two of the most important factors for understanding a simile. They also conducted a study of 16 similes to analyze responses from participants to understand the positive and negative impression a simile conveys toward the comparison subject, and concluded that difficulty in understanding a simile contributes to vagueness in interpreting sentiment in a simile. Hanks (2005) analyzed nouns that are used as the objects of comparison in similes. They were manually categorized into semantic categories (e.g., animals, roles in society, artifacts, etc.) that people most commonly use to compare with a subject. The use of animals (e.g., dog, fox) and mythical entities (e.g., dragon, angel) in metaphors has been analyzed by Rumbell et al. (2008) and Wallington et al. (2011).

2.3.2 Computational Models

Computational models for similes have also been explored in recent years. Veale and Hao (2007) extracted salient properties of concepts from the web using the "as ADJ as a/an NOUN" extraction pattern to acquire knowledge for the concept categories. Veale (2012) built a knowledge base of affective stereotypes by collectively analyzing all salient properties associated with the objects of comparison in similes. Li et al. (2012) used "as ADJ as" pattern as a simile template to query Google for similes, and determined the sentiment that properties convey toward famous persons, products, and companies across multiple lan-

guages. Niculae and Yaneva (2013) and Niculae (2013) used constituency and dependency parsing-based techniques to identify similes in text. Niculae and Danescu-Niculescu-Mizil (2014) designed a classifier with domain-specific, domain-agnostic, and metaphor-inspired features to determine when comparisons are figurative. They created a simile data set from Amazon product reviews, and showed that sentiment and figurative comparisons are correlated.

Computational methods for work on figurative language also include figurative language identification using word sense disambiguation, and determining sentiment polarity at the sense level using ngram graph similarity (Rentoumi et al., 2009), harvesting metaphors by using noun and verb clustering-based techniques (Shutova et al., 2010), interpreting metaphors by generating literal paraphrases in similar context using a framework for generating paraphrase from metaphors (Shutova, 2010), etc. More recently, the SemEval-2015 Shared Task 11 (Ghosh et al., 2015) has addressed the sentiment analysis of figurative language such as irony, metaphor, and sarcasm in Twitter. Their annotated data set contains 8,000 training tweets and 4,000 test tweets where the tweets are associated with an 11-point sentiment scale ranging from -5 to +5. Although their data set is expected to have irony, metaphor, and sarcasm, these categories (and any similes) are not explicitly identified or labeled in the data set. A comprehensive discussion on computational models of similes and metaphors can be found in the book by Veale et al. (2016).

In affective polarity recognition, one of the major differences between the work presented in this research and previous work is that this research is focused on determining affective polarity of a simile as a whole, where the affective polarity typically relates to an act or state of the comparison subject. Also, previous research most commonly used patterns to extract explicit properties from similes for various tasks, but none has addressed the task of inferring implicit properties in similes. Automatically inferring implicit properties in *open* similes and demonstrating their usefulness in affective polarity recognition is a novel contribution of the research presented in this dissertation.

2.4 Semantic Knowledge Bases and Lexicons

There has also been work to create more general semantic knowledge resources and lexicons. Contrary to sentiment or emotion lexicons, semantic lexicons typically map words and phrases to their general semantic category or contain relations such as hypernym or hyponym. Some semantic knowledge resources contain vast semantic knowledge, including word glosses, open or closed relations among entities, frames that indicate semantic roles of

entities in sentences, etc. The following sections discuss some of these knowledge resources.

2.4.1 Semantic Knowledge Bases

Some knowledge bases are manually created and contain term glosses, senses, and different types of semantic relations such as hypernym, hyponym, troponym, meronym, or antonym relations (e.g., WordNet (Miller, 1995)). Some are crowdsourced (e.g., DBpedia (Lehmann et al., 2015)) and in addition to semantic knowledge, can also contain commonsense knowledge (e.g., ConceptNet 5 (Speer and Havasi, 2013), Learner (Chklovski, 2003)). On the other hand, some systems are automatically built from the vast information in the web using information extraction techniques, and contain a more diverse set of entity relations (e.g., KnowItAll (Etzioni et al., 2004), NELL (Carlson et al., 2010), YAGO3 (Mahdisoltani et al., 2014), Probase (Wu et al., 2012)).

Although these knowledge bases are mainly created to represent general semantic knowledge of the world, they can contain semantic relations that are directly associated with people's affective states. For example, NELL (Carlson et al., 2010) contains relations such as diseases associated with emotions (e.g., autism is associated with *grief*, cancer is associated with *anxiety*) and plants representing emotion (e.g., daffodils represent *hope*, red roses represent *desire*). ConceptNet 5 (Speer and Havasi, 2013) contains both semantic and commonsense knowledge and has relations that describe more complex concepts causing affective states (e.g., "meet friend" or "win baseball games" cause *happiness*).

These knowledge bases are mainly created as repositories of general semantic knowledge. The affective knowledge that they contain is not explicitly learned as affective knowledge; rather they contain general semantic relations such as *associated_with(X,Y)* or *causes (X,Y)*. Sometimes arguments of these general relations happen to contain emotion or sentiment words, so the affective knowledge that can be acquired by exploiting these general relations is not as rich as traditional sentiment or emotion lexicons, and it is not trivial to identify the affective knowledge in these resources.

2.4.2 Bootstrapped Learning of Semantic Lexicons

Semantic lexicons are dictionaries that associate a word or term with its general semantic category (e.g., "cat" is an "animal"). Bootstrapped learning has been used to create lexicons so that only minimal human effort is needed. Many of these methods start with a few seed words for a semantic category and iteratively add new terms to the learned lists. For discovering candidate terms, these methods have considered nouns that appeared near seeds (Riloff and Shepherd, 1997) or utilized compound nouns and other syntactic constructions

(Roark and Charniak, 1998). Some work exploited syntactic heuristics (Phillips and Riloff, 2002), lexico-syntactic patterns (Riloff and Jones, 1999; Thelen and Riloff, 2002), weighted context N-grams (Murphy and Curran, 2007; McIntosh and Curran, 2008), predesigned and automatically learned context patterns (Pasca, 2004), domain-specific extraction patterns (Etzioni et al., 2005), doubly anchored hyponym patterns (Kozareva et al., 2008), and extraction patterns that are semantically related to the target categories (Qadir et al., 2015).

One of the challenges of iterative learning methods is that noisy inclusion of new category members affects successive iterations and may result in semantic drift. Thelen and Riloff (2002) learned multiple categories simultaneously to restrict the candidate term space of each category. Murphy and Curran (2007) used mutual exclusion bootstrapping to minimize semantic drift for both terms and contexts. McIntosh and Curran (2009) reduced semantic drift with bagging and distributional similarity. McIntosh (2010) learned negative categories when semantic drift has occurred. Carlson et al. (2009) simultaneously learned classifiers constrained with predefined entity relations. Qadir and Riloff (2012) designed an ensemble of component learning systems to learn only the category members that have consensus across different components.

In this work, emotion indicators are learned using a similar bootstrapping algorithm that starts with small number of seed terms per lexicon and iteratively grows the lists into bigger lexicons. But the lexicons contain emotion indicators that are hashtags and hashtag patterns, instead of words for general semantic categories. Unlike previous works that most commonly use lexico-syntactic or n-gram context patterns that are directly attached to seed words and target words, the method presented in this work trains n-gram classifiers with tweets that contain seed hashtags, and applies the trained classifiers to other tweets to iteratively learn new emotion hashtags and hashtag patterns.

2.5 Chapter Summary

In this chapter, I discussed previous work in the areas related to the research presented in this dissertation. In summary, the hashtag indicators of emotions addressed in this research are different from the words and phrases that can be found in traditional sentiment or emotion lexicons. They are fundamentally different types of tokens that are specific to social media text, and they present novel types of affective knowledge that have not been explored in the past by previous research. The emotion phrases harvested from the hashtags and the patterns can represent longer phrases beyond unigrams and bigrams typically found

in traditional sentiment or emotion lexicons.

The emotion indicators are learned using a similar bootstrapping algorithm as used by the traditional semantic lexicon induction techniques that begin with a small number of seed terms per category and iteratively grow the lexicons. But instead of lexico-syntactic or N-gram context patterns to extract category terms, in this research, emotion classifiers are trained using N-gram features during the bootstrapping iterations. The emotion classifiers are used to score, rank, and learn new emotion hashtags and hashtag patterns as emotion indicators.

A simile describes a state or activity of the subject of comparison, and the affective polarity of the state or activity is evoked from the comparison. While previous research presented methods to automatically learn affective events, the activities represented in similes typically have affective polarity in the contexts of the comparisons, so they offer a different type of affective knowledge about events. The research in this dissertation focuses on determining affective polarity of a simile as a whole, which is different from previous work on building a knowledge base of affective stereotypes or determining sentiment toward companies or famous people using similes as a useful linguistic device for these tasks. Automatically recognizing affective polarity in similes is a novel task addressed in this work.

None of the previous research works addressed the task of automatically inferring implicit properties in similes, which is fundamental for simile understanding. While previous research most commonly used explicit property patterns for various tasks, these tasks had different goals. Methods for automatically inferring implicit properties in *open* similes and using them to improve affective polarity recognition is a novel contribution of this research.

CHAPTER 3

BOOTSTRAPPED LEARNING OF EMOTION INDICATORS

People convey their affective states through various means. On the Twitter microblogging platform, people often use hashtags to express their emotional state (e.g., *#happyasalways*, *#angryattheworld*). Hashtags are created by adding the '#' symbol as a prefix to a word or concatenated multiword phrase without the whitespace. Knowing the emotion in a hashtag can be beneficial for understanding the overall emotion expressed by the writer in a tweet. In this research, I study hashtags as a source of affective knowledge.

To be able to use hashtags for recognizing people's affective state in a tweet, a repository of emotion hashtags is first needed that will allow one to know the emotion conveyed in a hashtag. As new hashtags are created by people everyday, manually creating such a repository by analyzing samples of tweets and the hashtags that appear in them will not be an effective solution for practical use. Fully supervised classification to automatically classify hashtags into emotion categories is possible, but manually acquiring labeled data is costly. A weakly supervised bootstrapping method can address this issues. The method would require a small number of seed emotion hashtags to jumpstart the learning process, so only little manual supervision will be required. The method can also be used to incrementally update the learned repositories of hashtags using new data.

In this chapter, I will present a bootstrapped learning framework that can be used to automatically learn a repository of emotion hashtags with respect to five emotion categories: Affection, Anger/Rage, Fear/Anxiety, Joy, and Sadness/Disappointment. The framework can be extended to also learn more general hashtag patterns, thus going beyond a list of specific hashtags in a repository. I will also discuss how emotion phrases can be harvested from the hashtags and hashtag patterns for contextual emotion classification. The automatic learning of emotion hashtags, emotion hashtag patterns, and the emotion phrases are the three types of hashtag indicators of emotions that I address in this research.

In Section 3.1, I will introduce the types of indicators that can be learned from Twitter hashtags as affective knowledge. In Section 3.2, I will discuss the selection of the emotion

categories for which the indicators are learned. In Section 3.3, I will present the bootstrapped learning framework for learning emotion hashtags and emotion hashtag patterns, and will discuss the design of a phrase context-based classifier that uses the emotion phrases harvested from the hashtags and the patterns. In Section 3.4, I will discuss the experimental set up for tweet emotion classification using the learned indicators and will describe the data sets. Finally, in Section 3.5, I will present the evaluation results and will discuss the impact of the learned hashtag indicators for emotion classification.

3.1 Hashtags, Hashtag Patterns, and Emotion Phrases

The first type of hashtag indicators of emotions that I address in this chapter are emotion hashtags. The hashtags that people use in tweets are often very creative. It is common to use single word hashtags (e.g., *#angry)*, but many hashtags are also multiword phrases (e.g., *#lovehimsomuch*). People often use elongated forms of words (e.g., *#yaaaaay*, *#goawaaay*) to put emphasis on their emotional state. In addition, words are often spelled creatively by replacing a word with a number or replacing some characters with characters that are phonetically similar (e.g., *#only4you*, *#Youredabest, #gr8*). They can also be shortened by removing some of the characters in a word (e.g., *#cantwait4tmrw*). These stylistic variations in the hashtags make it difficult to manually create a repository of emotion hashtags. For the same reasons, although emotion word lexicons exist (e.g., Mohammad, 2012a; Yang et al., 2007b), and hashtags can be artificially created by adding a '#' symbol as a prefix to the phrases in these lexicons, it will be unlikely to find all of the multiword phrases or stylistic variations that are frequently used in tweets.

Although some hashtags are common and used by many people repeatedly, many are also infrequent and very specific. One of the ways this can be addressed is by learning emotion hashtag patterns, which are the second type of hashtag indicators of emotions that I address in this chapter. I make the following observation that emotion hashtags often share a common prefix. For example, #angryattheworld and #angryatlife both have the prefix "angry at", which suggests the emotion ANGER. Consequently, these prefixes can be used to create hashtag patterns to generalize beyond specific hashtags and match all hashtags with the same prefix. For example, the pattern #angryat* will match both #angryattheworld and #angryatlife.

A key challenge is that a seemingly strong emotion word or phrase will sometimes not represent the writer's emotion. For example, #angry* may seem like an obvious pattern to identify ANGER tweets. But #angrybirds is another popular hashtag that refers to a

game, not the writers affective state. It is also possible that the emotion can be different depending upon the following words of a prefix phrase. For example, the phrase "love you" usually expresses AFFECTION when it is followed by a person (e.g., #loveyoumom), but it may express JOY in other contexts (e.g., #loveyoulife).

Emotion phrases are the third type of indicator that I address in this chapter. The learned hashtags and patterns can be used to harvest emotion phrases. For example, if the hashtag #lovelife is associated with JOY, then the phrase "love life" can be extracted from the hashtag and can be used to recognize this affective state in the main text of tweets. However, unlike hashtags, which are mostly self-contained,[1] the words surrounding a phrase in a tweet must also be considered. For example, negation can toggle polarity (*"don't love life"* may suggest SADNESS, not JOY) and the aspectual context may indicate that no emotion is being expressed (e.g., *"I would love life if ..."*).

3.2 Selection of Emotion Categories

For this research, five emotion categories were selected: AFFECTION, ANGER/RAGE, FEAR/ANXIETY, JOY and SADNESS/DISAPPOINTMENT. These categories were selected after analyzing Parrott's emotion taxonomy (Parrott, 2001) and observing the frequency of these emotions in tweets. The selected categories were found to be frequent in tweets and had minimal overlap with each other, making them ideal for the purpose of learning the hashtag indicators.

Two of Parrott's primary emotion categories were taken directly: JOY and FEAR, and these emotions are represented by all of their secondary and tertiary emotions. The FEAR category was renamed as FEAR/ANXIETY so that both emotions are equally represented by this class, since sometimes they overlap with each other and it is not always possible to distinguish one from the other. In Parrott's taxonomy, both FEAR and ANXIETY appear as tertiary emotions under the same primary emotion FEAR, but under different secondary emotions HORROR and NERVOUSNESS, respectively.

Parrott's secondary emotion classes AFFECTION and LUST were merged to form the AFFECTION class, and Parrott's secondary emotion classes SADNESS and DISAPPOINTMENT were merged to form the SADNESS/DISAPPOINTMENT class since these emotions are also difficult to distinguish from each other. Lastly, Parrott's secondary emotion RAGE was mapped to the ANGER/RAGE class. Both ANGER and RAGE appear in Parrott's taxonomy under the same primary emotion ANGER and secondary emotion RAGE. The primary

[1] Occasionally there can be a contextual effect (e.g., #thrilled #not).

emotion ANGER has other secondary emotions such as DISGUST or ENVY which were not used for this research. There were other emotions in Parrott's taxonomy such as SURPRISE, NEGLECT, etc., that were also not used in this research, mainly because 1) these other emotions are relatively less common in Twitter than the five selected emotions, and 2) some of them are difficult to recognize without larger context.

In addition to the five emotion categories, a NONE OF THE ABOVE class was used for tweets that do not carry any emotion or that carry an emotion other than one of the five emotion categories. Compared to the Ekman emotion classes (Ekman, 1992), one of the emotion taxonomies frequently used in NLP research (e.g., Mohammad, 2012b; Strapparava and Mihalcea, 2007), JOY, ANGER, SADNESS, and FEAR are comparable to four of the five selected emotion categories of this research. Ekman's SURPRISE and DISGUST classes were not studied in this work, but AFFECTION was additionally added.

3.3 The Bootstrapped Learning Framework

The hashtags and hashtag patterns are learned in a bootstrapped learning framework. The bootstrapping algorithm begins with a small collection of seed emotion hashtags, but iteratively grows the collection into a much bigger lexicon of hashtag indicators of emotions.

3.3.1 Overview

Figure 3.1 presents the framework of the bootstrapping algorithm for learning emotion hashtags and hashtag patterns. The bootstrapping process starts with five manually selected "seed" hashtags for each emotion category. These seed hashtags form the initial lexicon that is small in size. The end goal of the algorithm is to iteratively add more hashtags or patterns to the lexicon along with their mappings to the emotion categories.

For each seed hashtag in the initial lexicon, tweets that contain the hashtag are harvested

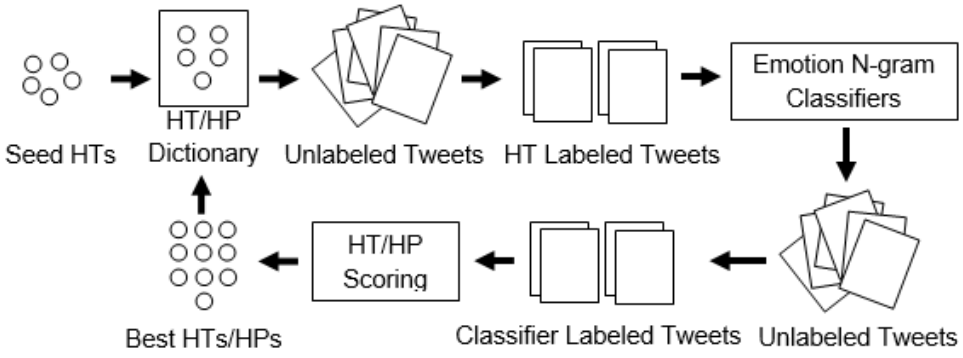

Figure 3.1: Bootstrapped learning framework (HT = hashtag; HP = hashtag pattern).

from Twitter. These tweets are then labeled with the corresponding emotion category. The labeled tweets are used to train a supervised classifier with N-gram features, for every emotion $e \in E$, where E is the set of emotion categories.

In the next step, the emotion classifiers are applied to a large pool of unlabeled tweets and the tweets that are labeled by the classifiers are retrieved. The hashtags that appear in these tweets are then scored and ranked. The most highly ranked hashtags are selected to add to the hashtag lexicon. Tweets in the pool of unlabeled tweets that contain the newly learned hashtags are labeled with the corresponding emotion and added to the set of training instances. The emotion classifiers are then retrained using the larger set of training instances, and the bootstrapping process continues.

To learn the more generalized emotion hashtag patterns, the same bootstrapping framework is used, but the hashtag patterns are learned separately from the hashtags because the candidate selection and scoring criteria are designed differently to account for multiple hashtags matched by a pattern.

3.3.2 Seeding

For each of the five emotion classes, five hashtags were manually selected as seed hashtags. To select the seed hashtags for an emotion class, first the adjectival forms of the emotions were identified which represented the corresponding affective states (e.g., Fear/Anxiety → afraid, anxious). Next, their close synonyms were identified (e.g., afraid → scared, angry → mad, pissed off, furious). Then hashtags were created by adding a '#' symbol as a prefix to these words and phrases after removing any whitespace. A few additional hashtags were directly identified from tweets that people commonly use to convey the corresponding emotion (e.g., #soulmate, #bff for the Affection class, #yay for the Joy class). Table 3.1 presents the seed hashtags.

These seed hashtags are strongly representative of the respective emotion. In the

Table 3.1: Seed emotion hashtags.

Emotion Classes	**Seed Hashtags**
Affection	*#loveyou*, *#sweetheart*, *#bff*, *#romantic*, *#soulmate*
Anger/Rage	*#angry*, *#mad*, *#hateyou*, *#pissedoff*, *#furious*
Fear/Anxiety	*#afraid*, *#petrified*, *#scared*, *#anxious*, *#worried*
Joy	*#happy*, *#excited*, *#yay*, *#blessed*, *#thrilled*
Sadness/Disappointment	*#sad*, *#depressed*, *#disappointed*, *#unhappy*, *#foreveralone*

selection of the seed hashtags, it was imperative that all of these hashtags are commonly used in Twitter, and that they do not belong to multiple emotion categories. The seed hashtags are then used to collect tweets that serve as initial training instances for a supervised classifier. To collect the tweets, Twitter is searched with the hashtags as the search keys, and the retrieved tweets are locally saved.

3.3.3 N-gram Emotion Tweet Classifiers

The tweets acquired using the seed hashtags serve as training instances to create emotion classifiers for supervised classification. First, the tweets were tokenized with CMU's freely available tokenizer for Twitter (Owoputi et al., 2013). Although it is not uncommon to express emotion in tweets with capitalized characters, the unique microblog writing styles often create many variations of the same words. Therefore, case normalization was done to ensure generalization. Also, re-tweets were removed to avoid repetition in training instances, and tweets with a URL were also removed because the emotional content could be in the external site for these cases.

Next, a logistic regression classifier was trained for each emotion class to predict a binary emotion label for a tweet. For the logistic regression algorithm, a freely available java version of the LIBLINEAR (Fan et al., 2008) package was used with its default parameter settings. Logistic regression was chosen because it produces probabilities with each prediction. These probabilities were later used to assign scores to candidate emotion hashtags. As features, unigrams and bigrams were used. Hashtags were also included as words in a tweet, except for the seed hashtags which were removed to avoid bias in the training data.

The expectation here is that the seed hashtag will not be the only emotion indicator in a tweet. The goal for the classifier is then to learn to recognize words and/or additional hashtags that are also indicative of the emotion. To avoid sparsity of the features, any N-gram that appeared only once in the training data and Twitter *usernames* (by looking for terms with '@' prefix) was removed.[2]

To train the classifier for emotion e, tweets containing the seed hashtags for e were used as the positive training instances and the tweets containing hashtags for the other emotions were used as negative instances. However, Twitter data also contain many tweets that do not have any emotion or express an emotion outside of the five targeted emotions. These tweets also need to be represented in the training data so that the classifiers can treat them

[2]A username is how someone is identified on Twitter, and is always preceded immediately by the @ symbol.

as negative training instances. For this purpose, 100,000 randomly collected tweets were added to the training data as additional negative instances. While it is possible that some of these tweets were actually positive instances for e, the expectation was that the vast majority of them would not belong to emotion e.

3.3.4 Learning Emotion Hashtags

The next step was to learn emotion hashtags. All of the trained classifiers were individually applied to all of the tweets in the unlabeled data to predict a binary label for each emotion (so a tweet could have multiple emotion labels). For each emotion e, the tweets classified as e were collected and the hashtags from those tweets were extracted to create a candidate pool H_e of hashtags for emotion e. To limit the number of candidates, hashtags that occurred < 10 times, had just one character, or had > 20 characters were discarded. Hashtags with just one character do not convey much meaning as emotion hashtags. Hashtags with many characters are often too specific. Removing these hashtags allowed for fewer hashtags to process during learning.

Next, each candidate hashtag h was scored by computing the average probability assigned by the logistic regression classifier for emotion e over all of the tweets (i.e., not just the tweets that are assigned emotion e) containing hashtag h. For each emotion class, the 10 hashtags with the highest scores were selected to add to the repository of learned hashtags. From the unlabeled tweets, all tweets with at least one of the newly learned hashtags were then added to the training instances, and the bootstrapping process continued.

Table 3.2 shows the top 20 hashtags that the bootstrapping process learned for the five emotion categories. The learned lexicons contain hashtags with creative spelling (e.g., #lonerlyfe, #singleprobz in SADNESS/DISAPPOINTMENT, #wuvyou, #youdabest in AFFECTION), acronyms (e.g., #tgfad in JOY which stands for "Thank God for Another Day"), and naturally occurring common spelling mistakes (e.g., #exicted in JOY which is a misspelling of #excited). The language recognizer used to identify English tweets did not always succeed, so a few internet meme hashtags from tweets written in other languages also got in (e.g., #countkun and #jadeinbekasi are hashtags that came from tweets written in Japanese and Indonesian). More of the top learned hashtags can be found in Appendix E.[3]

[3]The learned emotion hashtags can be downloaded from: `http://www.cs.utah.edu/~asheq/publications/data/emotion-indicators.zip`

Table 3.2: Top 20 hashtags learned using the bootstrapped learning model.

Affection	**Anger & Rage**	**Fear & Anxiety**	**Joy**	**Sadness & Disappointment**
#yourthebest	#godie	#hatespiders	#tripleblessed	#leftout
#notaprob	#donttalktome	#haunted	#tgfad	#foreverugly
#wishicouldbethere	#pieceofshit	#shittingmyself	#exicted	#singleprobs
#youdabest	#irritated	#worstfear	#thankful	#lonerlyfe
#otherhalf	#fuming	#scaresme	#24hours	#unloved
#youthebest	#hateliars	#nightmares	#birthdaycountdown	#jadeinbekasi
#bestfriendforever	#heated	#paranoid	#goodmood	#friendless
#flyhigh	#getoutofmylife	#hateneedles	#godisgood	#lonely
#loveyoulots	#angrytweet	#frightened	#greatmood	#teamlonely
#alwaysthere	#dontbotherme	#freakedout	#atlast	#heartbroken
#myotherhalf	#raging	#creepedout	#feelinggood	#notloved
#comehomesoon	#stupidbitch	#biggestfear	#happygirl	#singleprobz
#wuvyou	#madtweet	#sonervous	#godisgreat	#ineedfriends
#followher	#countkun	#shittingbricks	#lovemyfamily	#singleproblems
#alwaysandforever	#yourgross	#socreepy	#superhappy	#lonley
#alwayshere	#livid	#terrified	#newclothes	#needalife
#bestie	#screwyou	#waitinggame	#tentour	#lonertweet
#realfriend	#yousuck	#creeped	#newhair	#crushed
#missyousomuch	#badmood	#wimp	#liein	#miserable
#swimfast	#wankers	#nervous	#ecstatic	#letdown

3.3.5 Learning Hashtag Patterns

The hashtag patterns were learned using the same bootstrapping process but were learned separately from the hashtags. First, each hashtag was expanded into a sequence of words using an N-gram-based word segmentation algorithm.[4] The algorithm was supplied with corpus statistics from the Twitter corpus. The statistics contained unigram and bigram frequencies of all unigrams and bigrams of the corpus. For example, *#angryatlife* expands to the phrase *"angry at life"*. In a random sample of 100 hashtags, I estimated that expansion accuracy was 76% (+8% partially correct expansions).

All possible prefixes of the expanded hashtag phrases were then stored in a Prefix Tree (Trie)-like data structure. The prefixes consisted of words instead of characters, and were represented by nodes instead of edges. Next, the tries were traversed and all possible prefixes were accumulated from the nodes (excluding the terminal nodes) by visiting each path from root until a terminal node. The resulting path represented a candidate hashtag pattern. Exclusion of the terminal nodes ensured that the prefixes occurred with at least one following word. For example, *#angryashell*, *#angryasalways*, *#angrybird*, *#angryatlife*, *#angryatyou* would produce patterns: *#angry**, *#angryas**, *#angryat** as shown in Figure 3.2.

[4] `http://norvig.com/ngrams/`

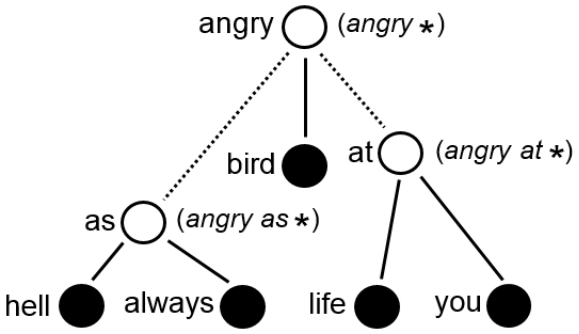

Figure 3.2: Candidate hashtag patterns represented in a Prefix Tree (Trie)-like data structure. Dotted lines lead to nonterminal nodes where patterns are extracted.

Each pattern was scored by applying the classifier for emotion e to all tweets having hashtags that matched the pattern. The classifier provided a binary decision for emotion e associated with the corresponding prediction probability for each tweet. The probability values were taken as scores and averaged for the pattern. For each emotion class, the 10 hashtag patterns with the highest scores were selected. From the *unlabeled tweets*, all tweets having hashtags that matched one of the learned hashtag patterns were then added to the training instances, and the bootstrapping process continued.

Table 3.3 shows examples of the learned hashtag patterns and matching hashtags. The matched hashtags of the patterns occasionally contained creative spelling (e.g., #bestiefolyfe) or acronyms (#bummedaf). This is one of the advantages of learning emotion hashtag patterns as the patterns will match other spelling variations too (e.g., #bestie* will match #bestiefolyfe, #bestieforlife, #bestieforlyfe, #bestie4lyfe, #bestie4life, etc.). More of the

Table 3.3: Examples of the learned hashtag patterns and matching hashtags.

Emotion	**Hashtag Pattern**	**Examples of Matching Hashtags**
Affection	#bestie*	#bestiefolyfe, #bestienight, #bestielove
	#missedyou*	#missedyoutoomuch, #missedyouguys, #missedyoubabies
Anger & Rage	#godie*	#godieoldman, #godieyou, #godieinahole
	#pissedoff*	#pissedofffather, #pissedoffnow, #pissedoffmood
Fear & Anxiety	#tooscared*	#tooscaredtogoalone, #tooscaredformama, #tooscaredtomove
	#nightmares*	#nightmaresfordays, #nightmaresforlife, #nightmarestonight
Joy	#feelinggood*	#feelinggoodnow, #feelinggoodforme, #feelinggoodabout
	#goodmood*	#goodmooditsgameday, #goodmoodmode, #goodmoodnight
Sadness & Disappoint.	#bummed*	#bummedout, #bummedaf, #bummednow
	#singlelife*	#singlelifeblows, #singlelifeforme, #singlelifesucks

learned emotion hashtag patterns can be found in Appendix E.[5]

3.3.6 Creating Phrase-based Classifiers

The third type of emotion indicator that I address in this research are emotion phrases. At the end of the bootstrapping process, the word segmentation algorithm was applied to all of the learned hashtags and hashtag patterns to expand them into phrases (e.g., *#lovemylife* → *"love my life"*). Each phrase can be assumed to express the same emotion as the original hashtag. However, unlike hashtags which are self contained and have topical focus, phrases are often influenced by the context in which they appear. For example, hypothetical situations or future actions may not convey the writer's emotion at the present time.

Consequently, a new logistic regression classifier was trained for each emotion e, which classifies a tweet with respect to emotion e based on the presence of a learned phrase for e as well as a context window of size six around the phrase (three words immediately on its left and three words immediately on its right). Tweets containing a learned phrase for e and a seed hashtag for e were the positive training instances. Tweets containing a learned phrase for e and a seed hashtag for a different emotion were used as the negative training instances.

I experimented with two variations of the context words feature. In the rigid context feature model, the positions of the context words were also taken into account. So for each word, there would be six features denoting if the word appeared n words before or after the phrase, where n can range from 1 to 3. In the flexible context feature model, two features (in place of six) denote if a word appeared before or after the emotion phrase regardless of its specific position.

For example, if *"love my life"* was learned as an emotion phrase for JOY, the tweet *"how can I love my life when everybody leaves me! #sad"* had one feature each for the left words *"how"*, *"can"*, and *"I"*, one feature each for the right words *"when"*, *"everybody"*, and *"leaves"*, and one feature for the phrase *"love my life"* under the flexible context feature model. Under the rigid context feature model, each of these words was a different feature, totalling six features for this instance. The tweet was then considered a negative instance for JOY because *"#sad"* (a seed hashtag for SADNESS/DISAPPOINTMENT) indicated a different emotion.

[5]The learned emotion hashtag patterns can be downloaded from: `http://www.cs.utah.edu/~asheq/publications/data/emotion-indicators.zip`

3.4 Data Sets in the Experimental Setup

3.4.1 Data Collection for Bootstrapping

The initial training data for the bootstrapped learning were collected by searching in Twitter for the seed hashtags shown in Table 3.1, using Twitter's Search API.[6] During the time of this data collection, tweets did not have a language meta-field to indicate the language in which the tweet was written. To ensure that the collected tweets were indeed written in English, a freely available language recognizer trained for tweets (Carter et al., 2013) was used. After filtering re-tweets and tweets with a URL, the seed labeled training data set contained 325,343 tweets. The distribution of emotions in this initial training data is presented in Table 3.4.

In addition to the seed labeled data, random tweets were collected using Twitter's Streaming API[7] to use as the pool of unlabeled tweets. Like the training data, re-tweets and tweets containing a URL as well as tweets containing any of the seed hashtags were filtered out. Since the research focus is on learning emotion hashtags, any tweet that did not have at least one hashtag was also filtered out. After the filtering steps, the unlabeled tweets collection contained roughly 2.3 million tweets.

3.4.2 Test Data

To create a gold standard data set, additional tweets separate from any of the learning or training tweets collection were acquired from Twitter, and were annotated by human annotators with respect to the five emotion categories or None of the Above.

Since manual annotation is time consuming, to ensure that many tweets in the test data had at least one of the five emotions, 25 topic keywords/phrases were manually selected that

[6] `https://dev.twitter.com/docs/api/1/get/search`

[7] `https://dev.twitter.com/docs/streaming-apis`

Table 3.4: Distribution of emotions in seed hashtag labeled and evaluation data sets.

Emotion	**Tweets with Seed Hashtags**	**Evaluation Tweets**
Affection	14.38%	6.42%
Anger/Rage	14.01%	8.91%
Fear/Anxiety	11.42%	13.16%
Joy	37.47%	22.33%
Sadness/Disappointment	23.69%	12.45%
None of the Above	-	42.38%

were considered to be strongly associated with emotions, but not necessarily any specific emotion. Then the topic phrases and their corresponding hashtags (i.e., hashtags created by adding a '#' symbol as a prefix to a topic phrase after removing any whitespace) were searched in Twitter to collect tweets that contained them. These 25 topic phrases were: *Prom, Exam, Graduation, Marriage, Divorce, Husband, Wife, Boyfriend, Girlfriend, Job, Hire, Laid Off, Retirement, Win, Lose, Accident, Failure, Success, Spider, Loud Noise, Chest Pain, Storm, Home Alone, No Sleep*, and *Interview*. This data collection process is similar to the emotion tweet data set creation by (Roberts et al., 2012). Since the purpose of collecting the tweets was to evaluate the quality and coverage of learned emotion hashtags, tweets not having at least one hashtag other than the topic hashtag were filtered out.

To annotate tweets with respect to an emotion, two annotators were given definitions of the five emotion classes from Collins English Dictionary,[8] Parrott's (Parrott, 2001) emotion taxonomy of these five emotions, and additional annotation guidelines (the annotation guideline can be found in Appendix A). A tweet can contain more than one emotion, or no emotion at all, so the annotators were instructed to label each tweet with up to two emotions or *None* indicating no emotion or a different emotion.[9] The instructions specified that the emotion must be felt by the writer.

The annotators reached an agreement level of 0.79 Kappa (κ) (Carletta, 1996) in an initial batch of 500 tweets. The annotation disagreements in these 500 tweets were then adjudicated by the two annotators, and each annotator labeled an additional 2,500 tweets. This produced an emotion annotated data set of 5,500 tweets. Example of the annotated tweets in different categories can be found in Appendix D.

1,000 randomly selected tweets from the collection were kept aside as a development/tuning data set, and the remaining 4,500 tweets were used as the evaluation data.[10] Table 3.4 shows the emotion distributions in the data sets. The distribution of tweets labeled using the seed hashtags is in the first column and the distribution in the test data is shown in the second column.

---

[8] `http://www.collinsdictionary.com/`

[9] It is possible for a tweet to contain more than two emotions, but these cases are rare.

[10] The evaluation data set can be downloaded from: `http://www.cs.utah.edu/~asheq/publications/data/Tweet-Emotion-Annotations.zip`

3.5 Evaluation of Emotion Indicators

The usefulness of the learned emotion indicators was determined in a tweet emotion classification task. The goal of the task was to predict the emotions felt by the writer in a tweet. To use the learned emotion hashtags and patterns for emotion classification in tweets, a lookup method was used. This method assigns an emotion category label to a tweet if the tweet contains a hashtag present in one of the learned emotion hashtag lexicons, or a hashtag that matches one of the learned prefix hashtag patterns. The classification decisions were binary for each emotion category, and a tweet could be assigned to multiple emotion categories.

The bootstrapped learning algorithm did not have a set stopping criteria and was run for 100 iterations. Some of the emotion categories were more prolific than others and learned substantially more emotion hashtags. The average probabilities that were used as scores for ranking and selecting new hashtags also varied across the emotion categories, so a threshold value could not be used consistently across all emotion categories to determine how many hashtags should be chosen.

To decide on the optimum size of the lexicons for each emotion class, instead, lexicon lookup was performed on the tuning data that were previously set aside before evaluation. For any hashtag in a tweet in the tuning dataset, the hashtag was looked up in the learned lexicons, and if found, the corresponding emotion was assigned as the label for that tweet. Starting with only seed hashtags in the lexicons, the sizes were incrementally increased by 10 hashtags at each trial. The optimum size was determined based on the best F-measure obtained for each emotion class. Table 3.5 shows the lexicon sizes that were found to achieve the best F-measure for each emotion class in the tuning data set.

Table 3.5: Optimum lexicon sizes decided from tuning data.

Emotion	**Number of Hashtags**	**Number of Hashtag Patterns**
Affection	260	390
Anger/Rage	940	810
Fear/Anxiety	920	970
Joy	440	270
Sadness/Disappointment	620	300

3.5.1 Baseline Systems

For comparison, a Support Vector Machine (SVM) classifier was trained with unigram and bigram features for each emotion class, using supervised classification with a 10-fold cross-validation setup on the evaluation data. I used the LIBSVM tool (Chang and Lin, 2011) and tuned the *cost* and *gamma* parameters for accuracy using the tuning data. The tuned parameter values were: $\gamma = 0.03125$ and c =8.

It is also possible that a hashtag repository can be trivially created by artificially adding a # symbol as a prefix to existing emotion lexicon phrases. To address this, I acquired the NRC Emotional Tweets Lexicon (Mohammad, 2012a), which contains emotion unigrams and bigrams for eight emotion classes. Four of the classes are comparable to the emotion categories addressed in this research: *Anger*, *Fear*, *Joy*, and *Sadness*. An artificial hashtag for each term in the lexicon was created by adding a # symbol as a prefix to a lexicon term. For bigrams, the two words in each term were concatenated without the whitespace. For N-grams that were associated with multiple emotion classes in the NRC Emotional Tweets Lexicon, the class with the highest score in the lexicon was chosen.

3.5.2 Evaluation of Hashtags and Patterns

The results for the N-gram classifiers and the lookup methods for classification using different lexicons are presented in Table 3.6. Two N-gram classifiers were used for classification, one using unigram features (SVM_1) and one using both unigram and bigram features (SVM_{1+2}). The table reports classification results using precision, recall, and F-measure for each emotion. The precision range is 63%-78%, but the recall of the N-gram classifiers is quite low, ranging from 10%–47%, leaving ample room for further improvement.

The middle section in Table 3.6 shows the emotion classification results using hashtags generated from the NRC N-grams Tweets Lexicon. Although the NRC Tweets Lexicon was constructed from tweets, the hashtags created from that resource have low precision ranging from 26%–39%. This could be due to many general words in the lexicon that can not independently act as emotion indicators when transformed into a hashtag (e.g., "candy" or "idea"). The recall is also low, ranging from 12%–18%, indicating that many hashtags in the evaluation data never appeared in the NRC emotion lexicon as words or phrases.

Next, the lower section of Table 3.6 shows the results for the seed hashtags that were used to jumpstart the bootstrapping, and the hashtags (HTs) and hashtag patterns (HPs) learned during bootstrapping (separately and together). First, just the five seed hashtags are used to assess their coverage. As can be expected, the seed hashtags achieve high

Table 3.6: Emotion classification results for hashtag lexicons and patterns lookup (P = Precision, R = Recall, F = F-measure)

Evaluation	**Affection**			**Anger & Rage**			**Fear & Anxiety**			**Joy**			**Sadness & Disappoint.**		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
N-gram SVM Classifiers															
SVM_1	78	40	53	66	17	27	68	33	44	66	47	**55**	63	26	37
SVM_{1+2}	78	35	48	67	10	17	68	29	41	65	43	52	63	21	32
Hashtags from NRC Emotion Lexicon															
NRC Tweets Lexicon	n/a			26	16	20	39	12	18	36	13	19	28	18	22
Hashtag and Hashtag Pattern Lexicons from Bootstrapping															
Seed Hashtags	94	06	11	75	01	03	100	06	11	93	04	08	81	02	05
Hashtags (HTs)	82	34	48	63	23	34	60	37	46	81	13	22	72	28	40
Hashtag Patterns (HPs)	76	48	59	60	22	32	57	42	48	84	09	16	73	16	26
All HTs+HPs	74	51	**60**	56	27	**36**	55	47	**51**	80	15	25	70	29	**41**

precision but very low recall. The maximum recall observed is only 6% for AFFECTION and FEAR/ANXIETY.

The next two rows show the results for hashtags (HTs) and hashtag patterns (HPs) (the lexicons include the seed hashtags too). The hashtags achieve performance similar to the supervised SVMs, except JOY. For the learned hashtag patterns, recall improves by +14% for AFFECTION and by +5% for FEAR/ANXIETY, which illustrates the benefit of more general hashtag patterns. Recall did not improve for ANGER/RAGE, JOY, and SADNESS/DISAPPOINTMENT. Table 3.5 shows that the optimum number of hashtag patterns for these classes was less than the optimum number of hashtags, which could be a potential reason for the lower recall.

When the hashtags and hashtag patterns are combined (All HTs+HPs), recall improves by as much as +17% in AFFECTION and +10% in FEAR/ANXIETY over the use of just the hashtags (HTs row), and improves F-scores across the board compared to SVM_1, with the exception of the JOY class. For the JOY class, recall is much lower than the recall for SVM_1, although precision is high. Table 3.4 earlier showed that among the five emotion categories, tweets with JOY were the majority (22.33% of the evaluation tweets). But Table 3.5 showed that the optimum number of hashtag patterns for JOY was less than the other emotion classes, and optimum number of hashtags learned for JOY was second to lowest. Many of the JOY hashtags and patterns learned at later iterations were topical (e.g., #newyear, #travel, #fishing, #flipflops*, #summer*) that may also be found in tweets with other emotions than JOY or in tweets with no emotion. This could be a potential reason for a low optimum number of hashtags and patterns, resulting in an overall low recall for the JOY

class compared to the rest of the emotion categories.

3.5.3 Evaluation of Emotion Phrases

To evaluate the learned emotion phrases, lexicon lookup was performed for the phrases. If a phrase belonging to an emotion lexicon was present in an evaluation tweet, the tweet was assigned the corresponding emotion. For comparison, tweets were also labeled with an emotion by performing lexicon lookup using the NRC Tweets Lexicon.

The first two sections in Table 3.7 show that emotion phrase lookup performs poorly, both for the NRC phrases and for the emotion phrases learned from the hashtags and hashtag patterns, suggesting that labeling a tweet based solely on the presence of a phrase is not very accurate. The last two rows of Table 3.7 show the results for when the trained phrase-based classifiers are applied. The phrase-based classifiers (PC) yield higher precision, albeit with low recall. The Flexible Context (FC) model performed slightly better than the Rigid Context (RC) model for the Affection, Fear/Anxiety, and Joy emotion categories (with respect to F-measure), but results were generally similar.

3.5.4 Evaluation Using Hybrid Approach

Table 3.7 showed that the phrase-based classifiers were relatively weak when used for emotion classification directly. As the classifiers also output prediction probabilities with respect to the emotion classes, these prediction probabilities can be used as additional features to SVM_1. These features (using the Flexible Context Model) are referred as PC in Table 3.8.

The SVM_1 + PC row in Table 3.8 shows that adding five features (one for each emotion class with the probability of the phrase-based classifier as the feature value) resulted in

Table 3.7: Evaluation of emotion phrases (P = Precision, R = Recall, F = F-measure, HT=Hashtags, HP=HT patterns)

Evaluation	**Affection**			**Anger & Rage**			**Fear & Anxiety**			**Joy**			**Sadness & Disappoint.**		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
NRC Emotion Lexicon Lookup															
NRC Tweets Lexicon	n/a			13	40	20	15	23	18	29	54	37	17	42	24
Our Emotion Phrase Lookup															
HT Phrases	36	22	27	19	38	26	31	31	31	53	22	31	31	19	24
HP Phrases	38	27	31	34	18	24	33	36	34	62	13	22	35	06	11
Emotion Phrase Context Classifiers (PC)															
Rigid Context (RC)	58	06	11	53	07	12	60	17	26	67	11	19	51	06	11
Flexible Context (FC)	54	07	12	48	05	09	63	17	27	69	12	20	50	06	11

Table 3.8: Emotion classification result for hybrid approaches (P = Precision, R = Recall, F = F-measure, HT=Hashtags, HP=Hashtag Patterns, PC=probability feature from emotion phrase context classifier (Flexible Context Model)).

Evaluation	**Affection**			**Anger & Rage**			**Fear & Anxiety**			**Joy**			**Sadness & Disapp.**		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
SVM_1	78	40	53	66	17	27	68	33	44	66	47	55	63	26	37
SVM_1+PC	79	42	55	63	18	28	70	35	47	68	48	56	62	27	38
(SVM_1+PC) ∪ HT	75	55	63	59	35	44	59	55	**57**	68	53	**60**	62	43	**51**
(SVM_1+PC) ∪ HT ∪ HP	69	64	**66**	55	38	**45**	54	61	**57**	68	54	**60**	62	44	**51**

consistent 1-2 point recall and F-measure gains over the original SVM_1 baseline. This result suggests that the emotion phrases can provide additional information beyond bag-of-Ngrams.

Finally, the learned emotion indicators are used in a hybrid approach for emotion classification. The last two rows of Table 3.8 show the results with the hybrid system. In this method, a tweet is labeled with emotion e if EITHER the SVM_1 + PC labels it as e, OR the tweet contains a hashtag or hashtag pattern associated with e. When both hashtags and hashtag patterns are used alongside the classifier, this combined approach achieves substantially higher performance than any individual method across all five emotion classes, with improved F-scores ranging from +5% to +18% over the baseline classifiers. These results demonstrate that the different types of emotion indicators are complementary.

3.5.5 Summary of Results

Table 3.9 summarizes the findings by showing the macro-average results across the five emotion classes for the different approaches explored. Simply looking for a hashtag or hashtag pattern in the lexicons learned with bootstrapping yields performance comparable to a supervised classifier with unigram features. One of the advantages of using this method for emotion classification is that this method does not require any manually annotated training data for classifying emotions.

The best results are obtained through the hybrid approach that assigns an emotion label to a tweet if either a hashtag or hashtag pattern in the tweet is present in that emotion's lexicon, or if the emotion class is predicted by the SVM classifier that uses unigrams and the probability output of the phrase context-based classifier as features. Overall, the learned emotion indicators (hashtags, hashtag patterns, and emotion phrases) increased the F score from 44% to 57% over the SVM_1 baseline, due to a nearly 20% gain in recall.

Table 3.9: Macro averages across the five emotion classes (P = Precision, R = Recall, F = F-measure, HT=Hashtags, HP=Hashtag Patterns, PC=probability features from emotion phrase context classifier (Flexible Context Model)).

Evaluation	**Macro Average**		
	P	R	F
SVM Classifiers			
SVM_1	68	33	44
List Lookup			
HTs ∪ HPs	67	34	45
Hybrid Aproaches			
SVM_1 + PC	68	34	45
(SVM_1 + PC) ∪ HTs	65	48	55
(SVM_1 + PC) ∪ HTs ∪ HPs	62	52	57

3.6 Chapter Summary

In this chapter, I have presented research on automatically learning hashtag indicators of emotion from tweets. In summary, five emotion categories have been selected for study after collapsing Parrotts emotion taxonomy (Parrott, 2001): Affection, Anger/Rage, Fear/Anxiety, Joy, and Sadness/Disappointment. These emotions are common in Twitter. A None of the Above label represents no emotion or an emotion which is not one of these five emotions.

A bootstrapped learning framework is introduced to automatically learn emotion hashtags and hashtag patterns. Starting with a small collection of seed hashtags for each emotion category, supervised emotion classifiers with N-gram features are used to label tweets with emotions and in turn score and rank new hashtags to iteratively add to the repository. At the end of the bootstrapped learning, emotion phrases are harvested from the learned hashtags and hashtag patterns, and phrase-based classifiers are trained with the context words of the emotion phrases to determine when the emotion phrases can be reliably used as emotion indicators.

For bootstrapped learning, unlabeled data are collected using Twitter Search with seed emotion hashtags. The evaluation data set for tweet emotion classification is created by searching for 25 topic keywords and hashtags on Twitter. These topics are expected to have strong association with the emotion categories but not particularly with any specific emotion category.

Tweet emotion classification results show that the hashtags and the hashtag patterns improve emotion classification over supervised classification with N-gram features for most

of the emotion categories. The emotion phrases harvested from the hashtags and patterns are not always reliable by themselves, but training additional classifiers with the emotion phrases and their surrounding context provides added benefits to emotion classification in tweets. The best classification results were achieved using a hybrid approach that classifies a tweet based on both the prediction of a supervised classifier (with N-gram probabilities of the phrase-based emotion classifiers as features) and the emotion hashtags and hashtag patterns in a lexicon lookup method. The hybrid approach substantially improves emotion classification performance across all five emotion categories, improving macro average F-measure from 44% (SVM_1) to 57% (hybrid approach).

The evaluation demonstrates that hashtags, hashtag patterns, and emotion phrases can be successfully learned as affective state indicators. In social media platforms such as Twitter, hashtags commonly convey a writer's emotion but they are often creatively written making it difficult to create a repository manually. The bootstrapping algorithm presents a method for learning emotion hashtags automatically. Common prefixes from emotion hashtags can also be used to create hashtag pattens, allowing for to generalization beyond specific hashtags. From both emotion hashtags and emotion hashtag patterns, emotion phrases can be additionally harvested and used to recognize emotion in the body of a tweet.

CHAPTER 4

SIMILES AS A SOURCE OF AFFECTIVE KNOWLEDGE

A simile is a figurative comparison that typically describes a subject through the subject's likeness with another concept. *"Her face looks like a tomato"* is an example of a simile where the face's appearance is explicitly compared (using "like") with a tomato's appearance. A simile is also known as a predicative comparison as the predicate describes the subject of the comparison (Bredin, 1998). This description is typically a state description (e.g., *"face looks like a tomato"* describes the state of the face), or an activity description (e.g., *"Jane swims like a dolphin"* describes Jane's swimming).

The state or the activity described in a simile can have an affective polarity. The affective polarity can be expressed explicitly (e.g., *"Jane swims* **beautifully** *like a dolphin!"*), or can be evoked entirely from the comparison itself. For example, *"Jane swims like a dolphin"* is easily understood to be a compliment toward Jane's swimming because dolphins are known to be excellent swimmers. Knowing this affective polarity is valuable affective knowledge as it represents the writer's subjective perception about the simile's subject and the corresponding state/activity. In this research, I investigate similes as a source of affective knowledge.

In Section 4.1, I will first discuss how similes are used for making comparisons. In Section 4.2, I will discuss the factors that play a role in the overall affective polarity of a simile and the roles played by different simile components in both understanding the affective polarity and in realizing the property of comparison when the property is implicit. In Section 4.3, I will describe studies conducted to shed light on how common similes are in Twitter. In Section 4.4, I will describe creation of a data set where the similes are labeled with positive or negative affective polarity. Finally in Section 4.5, I will describe the steps taken to create a simile data set with implicit properties.

4.1 Making Comparisons with Similes

4.1.1 Definition and Compositional Form

In the literature, a simile has been defined as a figure of speech that compares two essentially unlike things, where the comparison is introduced by words such as "like" or "as" (Paul, 1970). It is an explicit comparison of entities which are normally not considered comparable, thus making the comparison figurative in some sense (Israel et al., 2004). For example, *"my lawyer is like a shark"* compares two dissimilar entities: "lawyer" and "shark" (Sam and Catrinel, 2006).

A typical simile consists of four key components: the **topic** or **tenor** (subject of the comparison), the **vehicle** (object of the comparison), the **event** (act or state), and a **comparator** (usually "as", "like", or "than") (Niculae and Danescu-Niculescu-Mizil, 2014). A **property** (shared attribute) can be optionally included as well (e.g., *"His face looks **red** like a tomato"*). The similes that have an explicitly mentioned property are known as *closed* similes, and the similes that do not have an explicit property are called *open* similes (Beardsley, 1981).

Although "like" and "as" are two of the most commonly recognized and acknowledged comparators that signal a simile, a simile can also be created in other ways. Israel et al. (2004) identified three essential properties that make up a simile. They are: 1) a simile is a comparison, 2) the comparison is explicit, and 3) the comparison is figurative. In ensuring the second, the explicit comparator is not restricted to only "like" and "as", and can also be created using other phrases. For example, *"The retirement of Yves Saint Laurent is the fashion equivalent of the breakup of the Beatles"* (Israel et al., 2004) and *"Her voice makes this song shine brighter than gold"* (Niculae and Danescu-Niculescu-Mizil, 2014) also satisfy these three conditions, and are examples of similes created using "equivalent of" and "than" as the comparators.

4.1.2 How Similes Differ from Metaphors

To discuss how similes are used for making comparisons, it is important to understand how metaphors and similes relate to and differ from each other. While they are both considered figurative comparisons, a metaphor is "an implied comparison between two dissimilar objects, such that the comparison results in aspects that normally apply to one object being transferred or carried over to the second object" (Sopory and Dillard, 2002). Aristotle's "comparison theory" considers metaphors as elliptical similes. For example, *"Sam is a pig"* is a metaphorical comparison which differs from the simile *"Sam is like a pig"* (Chiappe and Kennedy, 2000). The ellipsis occurs because "like" is absent in the

former, but present in the latter. A key distinction between the two forms of comparison is that in a metaphor, the comparison is implied or implicit, whereas in a simile, a comparator phrase explicitly marks the comparison.

In addition to the compositional difference, the two forms of comparison are also not always interchangeable (ODonoghue, 2009). For example, *"Mary sings like an angel"* will result in some information loss in the metaphor *"Mary is an angel"*, because the latter does not describe any particular action, or the metaphor may not be about singing at all. Creating an equivalent paraphrase of such a simile and making a metaphor that conveys the same meaning is not simple. The opposite transformation can also suffer the same for many metaphors. For example, *"he buried the idea"* has a metaphorical use of the word "bury" to indicate that the idea was dropped or let go. It is not straightforward to paraphrase the same metaphor with a simile under the ellipsis assumption where just adding a comparator phrase such as "like" will transform the metaphor into a simile.

Removing a comparator from a simile sometimes may not result in a metaphor, rather an assertive statement. For example, *"he is like a doctor"* signals a figurative comparison because it refers to some doctor-like properties in a person who, in all likelihood, is not a real doctor. But when the comparator is removed, the statement *"he is a doctor"* becomes a simple predicate nominal which will rarely have an intended metaphorical use, and will refer to an actual doctor in most cases.

The dissimilarity between the tenor and the vehicle in a comparison creates the potential for a subjective view toward the tenor. Regardless of whether a simile can be transformed into a metaphor, a simile can be expected to have explicit or implicit properties, which will be the basis of the comparison and a contributing factor in the affective polarity of the simile when such subjective view is present.

4.1.3 Figurativeness of the Comparison in a Simile

By definition, a simile is a figurative comparison. For some similes, it is easy to understand why the comparisons are figurative. For example, in *"Joe drives like a snail"*, we know that snails do not drive, rather the driving is being compared with a snail's movement, and the speed of a vehicle can not literally match that of a snail's. On the other hand, some comparisons are literal and should not be considered as similes. For example, *"Dan looks like my neighbor"* most likely literally compares the visual attributes of two people, and is a literal comparison. But there are many cases where this is not as straightforward. ODonoghue (2009) discussed the example *"Sam eats like a pig"* to illustrate that if Sam's eating is very similar to how a pig eats (i.e., by lowering his head toward the plate, spilling

crumbs and snorting), it raises the question if such comparison can be held as figurative, because in a strong sense, Sam would be literally eating like a pig. Addison (1993) discussed a similar issue with the closed simile *"Red as a rose is she"* to make the point that both rose and she can be literally red, and perhaps the difference in the degree of redness between the concepts is how the comparison can be treated as a figurative comparison.

ODonoghue (2009) also pointed out that when the vehicle and the tenor belong to two different semantic categories, this can be a potential indicator of a figurative comparison. As Sam is a person and a pig is an animal, a comparison between the actions of the two is therefore more figurative than literal. However, there are other cases where this explanation is not sufficient. For example, in similes: *"he talks like a politician"* or *"she acts like a celebrity"*, both the tenor and the vehicle fall under the "human" semantic category.

Weiner (1984) hypothesized that the distance between the two concepts in a taxonomic relation may offer an explanation. For example, in *"dogs are like wolves"* and *"penguins are like wolves"* (hypothetically considering the latter has an interpretation), the former would be more literal because both dogs and wolves are canines, whereas penguins and wolves will be more distant from each other in the taxonomy. But this explanation is still limiting in being able to differentiate figurative comparisons from literal, when it comes to cases like neighbor vs. politician or a celebrity.

A different theory by Ortony (1993) can explain these cases better. Ortony hypothesized that a comparison is literal when, with respect to the comparison, both the tenor and the vehicle have the same high salience property. But the comparison is figurative when one of these two components has high salience properties that do not apply to the other. In *"Dan looks like my neighbor"*, "Dan" and "neighbor" do not have any typical high salience property (e.g., snails are slow, tomatoes are red). The comparison likely compares their visual attributes (comparable to having high salience properties if the visual attributes are their distinctive visual features) that are common to both, making it a literal comparison. On the other hand, in a simile where the vehicle is a politician or a celebrity, they can be expected to have properties that are more salient for the vehicle (e.g., politicians make promises, talk diplomatically; celebrities are good-looking, fashionable, stylish) than they are for the subject of the comparisons, making it a figurative comparison.

However, ODonoghue (2009) argued against Ortony's example, *"Encyclopaedias are like gold mines"*. The most important salient property in this comparison is that both encyclopaedias and gold mines store something valuable, but this property has high salience that is common to both the tenor and the vehicle; rather, it is the actual difference in the

manner of how something valuable is stored in an encyclopedia and in a gold mine that should make the comparison a figurative one.

For affective understanding of similes, part of the goal of this research is to identify when a simile has an affective polarity regardless of the degree of figurativeness. In both *"Sam eats like a pig"* and *"Joe drives like a snail"*, the described activity by the tenor has a negative polarity regardless of whether Sam is literally eating like a pig or whether Joe's driving can literally be compared with a snail's movement. While these aspects are fundamental for understanding how people compare two concepts in a simile, this research does not address the influence of different degrees of figurativeness in a comparison on the overall affective polarity of a simile.

4.2 Factors Contributing to Affective Polarity

Multiple factors may contribute to the affective polarity of a simile. These factors include the information that individual simile components carry, whether the comparison as a whole evokes a subjective view of the tenor, etc. The following section discusses these factors.

4.2.1 Nonpolar Distinctive Characteristics

When a comparison is in the compositional form of a simile, a figurative comparison is not syntactically distinguishable from a literal comparison. There are dedicated research efforts that identify when comparisons are not literal but figurative (e.g., Niculae and Danescu-Niculescu-Mizil, 2014), but it is outside the scope of this research. For simplicity, all comparisons in simile constructions, either literal or figurative, will be referred to as similes.

Sometimes (especially in case of literal comparisons), a tenor is compared by some distinctive characteristics of the vehicle that do not necessarily represent a subjective view of the tenor. Rather they come from observations about the similarity of the two concepts. Table 4.1 presents examples of these cases.

In (a) and (b), bananas and sister are used as vehicles in the comparisons. Bananas in general are known to have a distinctive smell. When some other object gives off a similar odor, its smell can be compared with a banana's smell. But it does not necessarily

Table 4.1: Comparison examples where distinctive characteristics are nonpolar.

	Simile	**Polarity**
(a)	it smells like bananas	*neutral*
(b)	she looks like my sister	*neutral*

indicate a positive or negative state of the tenor. Similarly, if a person shares similar visual characteristics with another person's sister, a comparison between the two will not necessarily have an affective polarity. Thus, examples (a) and (b) are neutral in Table 4.1.

When these comparisons are presented in context, an affective polarity is still possible. For example, if the writer has a general liking/disliking toward the smell of a banana, or a negative/positive feeling toward the writer's sister, and if the similes are presented in such context, they can have affective polarity. But these cases will be very personalized and specific, and the polarity is not evoked by the comparison itself but by external context.

4.2.2 Component Word Polarity

A word in a simile component can have a positive or negative connotation or sentiment. This can determine the polarity a simile should ultimately have. Table 4.2 presents examples of these cases.

Examples (a) and (b) illustrate the cases where the overall polarity of a simile can be determined from the polarity of the event verb. In example (a), the verb "stink" directly indicates that the writer considers the smell bad, and the simile has a negative polarity. In example (b), the verb "love" is positive, resulting in an overall positive polarity for the simile.

A positive or negative word can be present as a premodifier of the vehicle. In example (c), by using the word "rotten", the writer indicates that the smell is bad. In example (d), the writer describes a positive quality of the person by mentioning the word "caring". Consequently, the similes have negative and positive polarities, respectively.

Examples (e) and (f) illustrate that the words in the vehicle component can have a positive or negative connotation. In example (e), the word "garbage" has a negative connotation and the smell of garbage is typically perceived as a bad smell. Thus, the simile

Table 4.2: Simile examples with polarity in component words.

	Simile	**Polarity**
(a)	it *stinks* like bananas	negative
(b)	she *loves* me like a sister	positive
(c)	it smells like *rotten* bananas	negative
(d)	she talks like a *caring* sister	positive
(e)	it smells like *garbage*	negative
(f)	it smells like *heaven*	positive
(g)	my *friend* is like a scientist	positive
(h)	my *enemy* is like a scientist	negative

has a negative polarity. On the other hand, the word "heaven" has a positive connotation, resulting in an overall positive polarity for the simile.

Examples (g) and (h) show that polarity in the tenor component can be reflected in the overall polarity of a simile. In example (g), the word "friend" has a positive connotation. The vehicle "scientist" is neither positive nor negative, but the overall comparison is a compliment to "friend", evoking an overall positive polarity. Example (h) has the same event and vehicle as example (g). But now with a tenor having a negative connotation, the overall polarity of the simile is negative in example (h).

4.2.3 Affective Polarity Evoked by Implicit Properties

One of the interesting aspects of similes is that positive or negative polarity can be evoked entirely from the comparison itself, despite not having any positive or negative words in any of the components. Furthermore, the implicit properties play a significant role in determining the resulting affective polarity, even though they are not necessarily positive or negative words themselves. We can typically identify these properties from world knowledge and our general understanding of the characteristics or salient properties of the compared concepts.

In Table 4.3, examples (a) and (b) have no positive or negative words; rather the affective polarity is understood from world knowledge regarding the property of comparison. Example (a) is positive because running like a horse indicates *fast* running or running *energetically* which is good, while example (b) is negative because turtles move *slowly*, and running slowly is not a good way of running. But neither horse nor turtle has a positive or negative connotation on its own.

Examples (c) and (d) illustrate that a prior connotation can even be overridden depending upon the property being compared. In general, the word "celebrity" tends to have a positive connotation and looking like a celebrity is generally a compliment because it means someone is *fashionable* or *good-looking*. But acting like a celebrity is a negative simile

Table 4.3: Simile examples with affective polarity.

	Simile	**Polarity**
(a)	runs like a horse	*positive*
(b)	runs like a turtle	*negative*
(c)	looks like a celebrity	*positive*
(d)	acts like a celebrity	*negative*
(e)	my phone feels like a feather	*positive*
(f)	my wallet feels like a feather	*negative*

because it alludes to negative attributes such as *narcissism* or *entitlement*.

Examples (e) and (f) present the cases where the polarity can be different based on the compared property. Example (e) is positive because it means the phone is *light* and does not weigh much, which is a desirable property for a phone, whereas (f) is generally negative because it suggests that the wallet is *light* and does not have much money.

The implicit properties play influential roles in deciding whether a simile will have a positive or negative polarity or no polarity at all. Thus, modeling these aspects can benefit automatic recognition of affective polarity in similes.

4.2.4 Role of Component Words in Implicit Property Inference

Unlike *closed* similes, the *open* similes do not explicitly mention the compared property. For them, the properties need to be interpreted by the reader. As illustrated in Section 4.2.3, understanding these implicit properties can be crucial for recognizing the affective polarity for many similes. There are two important aspects of inferring an implicit property: 1) finding a property that is related to at least one of the simile components, and 2) the property should semantically fit within the simile's intended comparison, thus restricting or constraining the properties.

It is well acknowledged in the literature that a compared property in a simile is often a salient property of the vehicle. The Glucksberg et al. (1997) property attribution model of metaphor comprehension theorizes that the properties originate from a vehicle. As a figurative comparison, the same also applies to a simile. The property is usually salient enough that the vehicle can be used to exemplify the property (Veale and Hao, 2007). Let's consider the closed simile, *"my bed is hard like a rock"*. Among the characteristics of a rock, the property "hard" will have strong salience. Even when the property is not explicitly mentioned in a simile, for example, in *"my bed is like a rock"*, it is easy to understand that the described state of the bed is that it is hard.

There are also many cases where the event component of a simile is semantically strong. As a result, properties can have a strong association with the event. Let's consider the example, *"he is buzzing like a fridge"*. The most likely properties for the simile are "humming" or "vibrating". From the vehicle fridge, salient characteristics or functions such as "cold temperature" or "preserves food" will not apply here. Although "humming" and "vibrating" can be also be characteristics of a fridge, they are more semantically tied to "buzzing" than "fridge".

Some verbs that are frequently used in similes are semantically much weaker. For example, a "to be" verb is too general to make any strong contribution by itself. In a

simile such as, *"he is like a robot"*, the verb "is" is not informative to help infer any specific property that will apply in this simile. The class of sensory verbs or the verbs of perception (e.g., look, sound, feel, taste, smell) are typically stronger than the "to be" verbs, but they are still semantically weaker than verbs like "buzz". For example, for *"my face looks like a tomato"*, a myriad of properties can be considered from the verb "look", such as, adorable, fashionable, dirty, clean, and perhaps also, red and round. But most of these properties will not apply in this context, rather only a few. So, the role these verbs play are more in restricting the space of possible properties. For example, when we think about the vehicle "tomato", some salient properties are red, round, soft, juicy, and ripe. Not all of them are commonly used with "look", but as we mentally process the simile taking into account the influence of the verb, we would confine our inference space to a smaller collection of properties that are also compatible with "look".

The tenor component can also play the role of restricting the property space. For example, in *"my face looks like a tomato"*, properties such as "ripe" or "juicy" are not compatible with "face", restricting the inference space. However, for some other similes, the role of the tenor is not as influential as the event component. For example, in "my room feels like Antarctica", most people will infer that the property is cold. Antarctica can also be associated with other salient properties such as white, beautiful, big, etc. While these other properties are not strongly associated with the verb "feel", they are still compatible with the tenor "room", and the impact the tenor will have in restricting the inference space is limited in cases like this. So, there is certainly evidence of both cases for when a tenor is influential and when it is not.

When it comes to generating a property from the tenor, theoretical models of figurative comparison suggest that properties are typically attributed to or applied to the tenor (Glucksberg et al., 1997), instead of being generated from the tenor. It is, however, certainly possible to find properties associated with some tenors when the tenors are semantically strong. For example, in *"my eyes feel like clams"*, some of the most fitting properties are: squinty, heavy, weary, which are strongly associated with "eye". Other times, a tenor can simply be pronouns or named entities, providing very little information with respect to generating a property. For example, in *"John drives like a snail"*, without context, a reader will not know anything else about John's characteristics. But this does not prevent the reader from understanding that the property in this simile is *slow*, despite not knowing who John is.

4.3 Study on How Common Similes Are in Twitter

To understand how common similes are in Twitter, I conducted analyses on multiple samples of English Tweets. As a popular microblogging platform, Twitter contains similes that people use in everyday messages. Twitter is also widely used for sentiment analysis, which makes it an ideal source for collecting similes with affective polarity. The Twitter corpus used in this research consists of roughly 140 million English tweets harvested using the Twitter streaming API from March 2013 to April 2014.

For this study, first a random sample of 10,000 tweets was drawn from the Twitter corpus. From the sample, tweets that contained the keyword "like" and the pattern "as X as" were extracted, where 'X' is a single word. The resulting set of tweets was then manually analyzed to identify if they contained a simile. A total of 68 similes were identified in the sample (0.68%), indicating that a simile can be found in nearly every 147 random English tweets. Among the identified similes, 94% appeared with the keyword "like" and the remaining 6% had the "as X as" construction.

The identified similes were further analyzed to estimate how often properties are explicitly mentioned in similes from Twitter, i.e., the relative frequency of *open* and *closed* similes. *Among the identified 68 similes, 54 similes (92%) did not have an explicitly mentioned property, which suggests that the vast majority of the similes in Twitter are* ***open*** *similes.*

Next, a second study was conducted to understand if similes are more common in tweets that contain a positive or negative sentiment. However, manually identifying tweets with a positive or negative sentiment from a large sample would have required a separate sentiment annotation task. As an alternative, I used emoticons as a proxy for the sentiment labels. 10,000 random tweets were drawn from the Twitter corpus that either contained a happy-face emoticon (":)" or ":-)") or a sad-face emoticon (":(" or ":-(") but not both, so that the sample represents tweets with either a positive or a negative sentiment. 7,641 out of the 10,000 tweets were labeled with positive sentiment as they contained the happy-face emoticon, and the remaining 2,359 tweets were labeled with negative sentiment as they contained the sad-face emoticon. As before, from the sample, tweets containing "like" and "as X as" were extracted and among them, tweets that contained a simile were manually identified.

Among the 7,641 positive sentiment tweets, 75 of them were identified to have a simile (0.98%). Among the 2359 negative sentiment tweets, 15 of them were identified to have a simile (0.64%). Combined, a total of 90 tweets out of 10,000 sentiment tweets contained a simile (0.90%), suggesting that a simile can be found in every 111 tweets that have a

positive or negative sentiment.

Emoticons are only rough approximations for positive/negative sentiments, so I also looked into manually annotated data that are widely used in sentiment analysis research. For this purpose, I used the SemEval 2013 Task 2 Subtask B training data (Nakov et al., 2013)) containing manually annotated tweets with respect to positive, negative, and neutral categories. However, it must be noted that this data set is not a random collection of English tweets. They were collected by searching for named entities that were popular topic keywords in specific time periods. When the data was reacquired from Twitter using the tweet id information, there was a total 6,425 tweets still available in Twitter (64% of the original data set), out of which, 2,343 were positive sentiment tweets, 904 were negative sentiment tweets, and 3,178 were neutral tweets.

Using the same method as before, tweets containing a simile in the SemEval data were identified. Among them, 16 tweets out of the 2,343 positive sentiment tweets were found to have a simile (0.68%), and 23 out of the 904 negative sentiment tweets were found to have a simile (2.54%). Combined, 39 out of 3,247 tweets with either positive or negative sentiment contained a simile, indicating that when tweets contain a popular named entity topic with some association to sentiment, a simile can be expected in every 111 tweets (0.90%) having either positive or negative sentiment.

In all of the studies, similes that contained a profanity word (e.g., *"I look like shit"*) or frozen expressions that are not figurative comparisons (e.g., *"it sounds like a good idea"*) were excluded from considerations when the similes were identified.

Table 4.4 summarizes the statistics of the studies. The findings suggest that similes are

Table 4.4: Statistics for how common similes are in tweets.

Tweets	**# of Tweets with Similes**	**Total Tweets**	**Percentage**	**Expected Simile Occurrence**
Random Twitter Stream				
Random	68	10,000	0.68%	1 in every 147 tweets
Emoticon Labeled Data				
Positive Sentiment	75	7,641	0.98%	1 in every 157 tweets
Negative Sentiment	15	2,359	0.64%	1 in every 102 tweets
Combined	90	10,000	0.90%	1 in every 111 tweets
SemEval 2013 Task 2 Subtask B Training Data				
Positive Sentiment	16	2,343	0.68%	1 in every 146 tweets
Negative Sentiment	23	904	2.54%	1 in every 39 tweets
Combined	39	3,247	1.20%	1 in every 83 tweets

more common in sentiment tweets than they are in random tweets. Also, the findings seem to suggest that similes are more common in tweets with a negative sentiment than they are in tweets with a positive sentiment.[1]

4.4 Creating Simile Data Sets with Affective Polarity

In this section, I will describe the creation of a simile data set where the similes are associated with affective polarity. To create the data set, similes are first extracted from the Twitter corpus mentioned in the previous section, using extraction patterns that conform to the syntax of similes. The similes are then assigned affective polarity category labels to be used as gold standard.

4.4.1 Simile Extraction and Data Preprocessing

As the first step of the data set creation, all tweets that contained any of the three commonly used **comparator** keywords: *"like"*, *"as"*, and *"than"* were selected. These three keywords were also used by Niculae and Danescu-Niculescu-Mizil (2014) for creating a data set of comparison statements.

Many of the tweets had exact duplicate content, in which case only one tweet was kept and the duplicates were discarded. A common phenomenon in Twitter is re-tweeting, where a writer posts someone else's tweet, either with exact content or with some modification. Re-tweets are problematic because they introduce lexical bias in the data set due to overlapping content. Re-tweets often contain a re-tweet token (e.g., "rt" or "#rt") that acts as a marker indicating that the tweet was not originally written by the person who posted it. From the tweet collection, any tweet that contained a re-tweet token was removed.

An additional challenge of a tweet corpus is near duplicate content. Most of the near duplicate tweets are not spontaneously created by different people; rather these tweets often contain a memorable or interesting quotation or song lyric. The tweets slightly differ because the writers change some of the words or add a few additional words typically at the beginning or end of a message. Table 4.5 presents some examples of near duplicate tweets containing a simile.

Examples (a)-(e) show that the variations in near duplicate tweets can occur due to the use of different words (e.g., "best" and "sweetest"), elongated words (e.g., "tonight" and "toooooonight"), spelling mistakes (e.g., "rhianaa" and "rihanna"), or additional words added at the end of a message (e.g., "- lillian dickson" and "#truth"). Examples (a), (c),

[1]The biases in the sentiment data collection methods should be taken into account.

Table 4.5: Example of similes in near duplicate tweets and their Jaccard similarity score for trigram overlaps. Tokens that are different in a pair of tweets are in boldface.

ID	Tweets	Jaccard Similarity
(a)	Tweet1: <s><s>**i mean !!!** : a relationship where you can act like complete idiots together is the **best** thing ever . ” </s></s> Tweet2: <s><s>**“** : a relationship where you can act like complete idiots together is the **sweetest** thing ever . ” </s></s>	0.52
(b)	Tweet1: <s><s>you ’ll be coming home with me **tooooonight** , and we ’ll be burning up like neon **liiiiiiiiiiiiights** ! </s></s> Tweet2: <s><s>you ’ll be coming home with me **tonight** , and we ’ll be burning up like neon **lights** </s></s>	0.52
(c)	Tweet1: <s><s>life is like a coin . you can spend it any way you wish , but you only spend it once . **- lillian dickson** </s></s> Tweet2: <s><s>**:** life is like a coin . you can spend it any way you wish , but you only spend it once . **#truth** </s></s>	0.61
(d)	Tweet1: <s><s>she ca n’t sing she ca n’t dance but who cares she walks like **rhianna** </s></s> Tweet2: <s><s>she ca n’t sing she ca n’t dance but who cares she walks like **rihanna #mtvstars beyonce** </s></s>	0.62
(e)	Tweet1: <s><s>when i miss you , i re-read our old conversations and smile like an idiot . </s></s> Tweet2: <s><s>when i miss you , i re-read our old conversations and smile like an idiot . **i miss u darl..@fitripitox** </s></s>	0.67

and (e) are quotations or internet memes, and examples (b) and (d) are lyrics from songs by singer Demi Lovato and the band The Wanted.

To identify these cases, a de-duplication step was performed using Jaccard similarity of trigrams to measure any overlap in the text content between each pair of tweets. Jaccard similarity was calculated by:

$$jaccard(A, B) = \frac{T_a \cap T_b}{T_a \cup T_b}$$

where T_a and T_b are the trigram sets from tweets A and B, respectively. Whenever Jaccard similarity was higher than 0.5, the larger tweet (in terms of words) was kept, and the process was repeated until no such pair with similarity higher than 0.5 remained.

After the de-duplication steps, the UIUC Chunker (Punyakanok and Roth, 2001) was used to identify phrase sequences representing the syntax of similes. Table 4.6 lists these phrase sequences. For example, when a tweet contains the phrase sequence: NP_1 + VP + ADJP + PP-like + NP_2, it represents a simile where NP_1 is the tenor, NP_2 is the vehicle, VP is the event, and ADJP is an explicitly mentioned property.

Table 4.6: Phrase sequences used to extract similes from tweets.

Phrase Sequence	Example
Comparator: like	
NP_1 + VP + PP_{like} + NP_2	my room feels like an oven
NP_1 + VP + ADJP + PP_{like} + NP_2	my room feels hot like an oven
NP_1 + VP + ADVP + PP_{like} + NP_2	he is acting immaturely like a child
NP_1 + VP + NP_2 + PP_{like} + NP_3	he tricked me like a conman
NP_1 + VP + NP_2 + ADVP + PP_{like} + NP_3	he tricked me cleverly like a conman
Comparator: than	
NP_1 + VP + ADJP + PP_{than} + NP_2	my room feels hotter than an oven
NP_1 + VP + ADVP + PP_{than} + NP_2	he is acting more immaturely than a child
NP_1 + VP + NP_2 + ADVP + PP_{than} + NP_3	he tricked me more cleverly than a conman
Comparator: as	
NP_1 + VP + $ADJP_{as}$ + PP_{as} + NP_2	my room feels as hot as an oven
NP_1 + VP + $ADVP_{as}$ + PP_{as} + NP_2	he is acting as immaturely as a child
NP_1 + VP + NP_2 + $ADVP_{as}$ + PP_{as} + NP_3	he tricked me as cleverly as a conman

The extracted similes were generalized by removing the comparator and the optional explicit property component, so that affective polarity can be recognized even when a property is not explicitly mentioned. For further generalization over the lexical forms, words in the similes were lemmatized using Stanford CoreNLP (Manning et al., 2014) and they were normalized with respect to case. For a tenor phrase, the head noun is usually sufficient to understand the affective polarity target, so only the head noun was kept. For vehicles, any leading article or determiner was removed, and the rest of the noun phrase was kept. This is because vehicles with different noun or adjective modifiers like "ice box" and "gift box" may represent two different concepts with different polarities in similes. Any pronoun that refers to a person was replaced with a general "PERSON" token and other pronouns with "IT". Table 4.7 lists these pronouns.

Sometimes, vehicle phrases contain adjective modifiers indicating a sentiment (e.g., *"she looks like a **beautiful** model"*). In these cases, the sentiment is explicit and external to the comparison. Additionally, similes sometimes contain profanity (e.g., *"You look like **crap**"*), which are typically negative, and have become commonly used expressions that are not intended to be figurative comparisons. The focus of this research is to recognize affective polarity in similes where the affective polarity is evoked from the comparison itself.

Table 4.7: List of pronouns replaced with PERSON and IT tokens in tenor of similes.

Replaced Token	Pronouns
PERSON	i, me, you, u, he, him, she, her, we, us, they, them, someone, anyone, who, whom, everyone
IT	it, what, this, that, these, those

Therefore, if a simile vehicle contained a sentiment-bearing adjective modifier, the simile was removed from the data set. These cases were identified using the AFINN sentiment lexicon (Nielsen, 2011), which is widely used for tweet sentiment classification. For identifying and filtering out similes containing profanity, a freely available list[2] of profanity words was used.

Any simile where the vehicle is a pronoun (e.g., *"it looks like **that**"*) was also removed, because without resolving the pronoun, the simile offers little to no information for understanding the comparison. To avoid infrequent cases, similes appearing fewer than 5 times were discarded. Each simile was then represented by a triple of the tenor, event and vehicle (e.g., *"my face looks red like a tomato"* → <face, look, tomato>). The remaining set after these filters were applied contained 7,594 similes.

4.4.2 Manual Annotation of Affective Polarity

To obtain manual annotation, 1500 similes with frequency ≥ 10 were selected out of the 7,594 similes, with the expectation that the more frequent similes will be easier for the human annotators to understand. Amazon's Mechanical Turk was used to obtain gold standard annotations for affective polarity. The annotators were asked to determine if a simile expresses affective polarity toward the subject (i.e., the **tenor** component) just by reading the simile in isolation, so that people's general perception or stereotypical views about them can be captured regardless of any context (although each simile was presented to the annotators along with three tweets to show examples of how the simile can be used because sometimes it is difficult to understand the meaning of a simile in isolation).

The annotators were asked to assign one of four labels: *positive*, *negative*, *neutral*, or *invalid*. The first two labels are for similes that clearly express positive polarity (e.g., *"Jane swims like a dolphin"*) or negative polarity (e.g., *"Fred's hair looks like a bird's nest"*). The *neutral* label is for comparisons that do not have polarity (e.g., *"Dan looks like my neighbor"* is not a positive/negative comment about Dan) or similes that are ambiguous without the benefit of context (e.g., *"he is like my dog"* could be good or bad depending on the context).

The data also contained many misidentified similes, typically due to parsing errors. For example, sometimes there is an entire clause in place of the vehicle (e.g., "I feel like I'm gonna puke"). Other times, the informal text of Twitter makes the tweet hard to parse (e.g., "he is like whatttt") or a mistagged verb occurs after "like" (e.g., "he is like hyperventilating"). The *invalid* label covers these types of erroneously extracted similes. Table 4.8 presents examples of positive and negative similes from the annotated data.

[2] `http://www.bannedwordlist.com/lists/swearWords.txt`

Table 4.8: Sample similes with positive/negative polarity from the annotated data.

Positive	**Negative**
<PERSON, smile, sun>	<PERSON, look, zombie>
<PERSON, feel, kid>	<PERSON, treat, stranger>
<PERSON, be, older brother>	<PERSON, feel, poo>
<IT, sound, heaven>	<PERSON, look, clown>
<PERSON, look, superman>	<word, cut, knife>
<IT, be, old time>	<PERSON, act, child>
<IT, feel, home>	<PERSON, look, voldemort>
<IT, fit, glove>	<PERSON, look, wet dog>
<IT, would be, dream>	<PERSON, treat, baby>
<IT, smell, spring>	<PERSON, look, drug addict>

The annotation task was first conducted on a smaller scale, with a set of 50 similes, to select workers who had high annotation agreement with each other and gold standard labels prepared by me. The best three workers then all annotated the official set of 1500 similes. The average Cohen's Kappa (κ) (Carletta, 1996) between each pair of annotators was 0.69. The final labels were determined through majority vote. However, none of the annotators agreed on the same label for 78 of the 1500 similes, and 303 instances were labeled as *invalid* similes by the annotators. These instances were subsequently excluded from the annotated data set. The remaining similes were randomly divided into an evaluation set (Eval) of 741 similes, and a development set (Dev) of 378 similes.

Table 4.9 shows the label distribution of the data sets. Overall, a very high percentage (89.37%) of the annotated similes (combining development and evaluation data) had affective polarity. The annotation guideline can be found in Appendix B, and more examples from the annotated data in Appendix F.[3]

Table 4.9: Distribution of labels in the manually annotated development and evaluation data sets.

Label	**# of Similes (Dev Data)**	**# of Similes (Eval Data)**
Positive	164	312
Negative	181	343
Neutral	33	86
Total	378	741

[3]The annotated data set can be downloaded from: `http://www.cs.utah.edu/~asheq/publications/data/simile-dataset.zip`

4.5 Creating a Simile Data Set for Implicit Property Inference

Explicit properties are important for understanding the meaning of a simile, and in turn can play an important role in recognizing the affective polarity in a simile. When they are not explicitly mentioned, they need to be inferred. The sample study of similes in Twitter described in Section 4.3 estimated that 92% of the similes are *open* similes that do not mention an explicit property. In this section, I describe the creation of a gold standard data set consisting of *open* similes with their implicit properties.

4.5.1 Collecting Similes with Implicit Properties

As the first step for creating the data set, similes that match the syntax of *open* similes were extracted from the Twitter corpus. A part-of-speech tagger designed for Twitter (Owoputi et al., 2013) was applied to tweets containing the word "like". The other two comparators, "as" and "than", that were used in the data set creation described in Section 4.4 were not used here because these two comparators are only used in *closed* similes.

Next, the UIUC Chunker (Punyakanok and Roth, 2001) was applied to recognize noun phrases and verb phrases in these tweets. Tweets were then selected that matched the syntactic pattern: $NP_1 + VERB + like + NP_2$, where NP_2 can contain only a noun and an optional indefinite article. The similes were required to have a vehicle term with no premodifier to avoid problems associated with coreference (e.g., "the man" or "that man") and to focus on vehicles that represented general concepts. It is possible to have vehicles which are multiword phrases or clauses (e.g., *"my room is like stepping into a hurricane"*, *"my room is like a boots store"*, *"my room looks like a tornado has hit it"*, etc.). These cases are not addressed in this research and are left for future work.

The selection process extracted many figurative similes, but it also extracted literal comparisons with no apparent property (e.g., *"this flower smells like a rose"*) and statements that were not comparisons (e.g., *"I called like five times"*). To focus on figurative similes with an implicit property, as a second step, the similes were further filtered for vehicle terms that occurred in comparisons with an explicit property. To identify vehicles that have been previously seen with explicit properties in the corpus, using the same Twitter data, 995 nouns were extracted that appeared in specific syntactic patterns. These syntactic patterns represent comparison constructions with an adjectival property: $ADJ + like + [a, an] + NOUN$ (e.g., *"red like a tomato"*) and $ADJ + as + [a, an] + NOUN$ (e.g., *"red as a tomato"*). In the simile collection, only similes whose vehicle is one of the 995 nouns

were kept. Finally, similes that contained pronouns, common person first names[4] (to avoid issues with coreference resolution), or profanity,[5] were filtered out. Similes with words not in a dictionary [6] were also filtered out to avoid issues with Twitter language such as misspellings, elongated words, etc. As before, the words in the similes were lemmatized using Stanford CoreNLP (Manning et al., 2014) and were lowercased. A total of 3,011 *open* similes discovered with frequency ≥ 3 were compiled.

4.5.2 Gold Standard Implicit Properties

Next, a gold standard set of implicit properties was generated for each simile using Amazon's Mechanical Turk. First, seven Mechanical Turk workers prequalified, and each annotated 700 similes. The 700 similes were randomly selected from the 3,011 *open* similes. Each annotator was asked to provide up to 2 properties that best captured the most likely basis for comparison between the tenor and vehicle.

The annotators were also provided with the option to label a simile as *Invalid* if it was not a simile at all (most commonly due to parse errors, such as *"he looks like ran"*) or label a simile as having No Property (often due to literal or underspecified comparisons, such as *"she looks like my aunt"*). The annotators were asked to give adjectives, adverbs, or verbs, but occasionally they provided a noun.

Among the 700 similes, a majority of the annotators labeled 59 of them as either *Invalid* or *No Property*. These similes were removed from the data set. The resulting 641 similes had 9.84 properties per simile on average. Out of the 641 similes, 183 similes (29%) were then set aside as a development set and the remaining 458 similes (71%) were kept as a test set. Table 4.10 presents examples of annotated properties in the gold standard. The annotation guideline can be found in Appendix C, and more examples from the annotated data in Appendix G.

4.6 Chapter Summary

In this chapter, I discussed similes as a source of affective knowledge. In summary, similes describe the state or activity of the subject of comparison. In an annotated gold standard data set, 89.37% of similes have been observed to have affective polarity.

---

[4] `http://deron.meranda.us/data/census-derived-all-first.txt`

[5] `http://www.bannedwordlist.com/lists/swearWords.txt`

[6] Using Wordnik: `https://www.wordnik.com/`

Table 4.10: Similes with sample properties inferred by human annotators.

Simile	**Properties Inferred by Humans**
<laugh, be, music>	melodic, pleasing, dulcet, tinkly, enjoyable
<PERSON, sound, prophet>	wise, insightful, prescient, enlightened, foreseeing
<eye, feel, clam>	slimy, squinty, weary, gummy, heavy
<PERSON, look, carrot>	orange, thin, scrawny, slim, tall
<PERSON, buzz, fridge>	humming, vibrating, distracting, annoying, motorized
<PERSON, fight, animal>	ferociously, scratches, tenaciously, wild, aggressive
<PERSON, be, shark>	sneaky, primordial, dangerous, predatory, opportunistic
<time, be, river>	flowing, fast, winding, unending, moving
<praise, be, sunlight>	warm, rejuvenating, energizing, cheerful, pleasing

As a figurative comparison, a simile has the following major components: tenor (subject of comparison), event (verb of a simile) and vehicle (object of comparison). A simile can optionally have a property component. The similes that have a property are known as *closed* similes, and when the property is implicit, they are known as *open* similes. Also, similes are different from metaphors because similes are explicit comparisons that must have a comparator, and one can not always be transformed into the other. By definition, the comparison is figurative in a simile, but the literal versus figurative distinction is not always straightforward.

Multiple factors can contribute to the affective polarity of similes. Positive/negative connotations or sentiments in simile component words can have an important influence in the overall affective polarity of the simile, but the polarity can also be evoked entirely from the comparison itself. Properties can be important for understanding the comparison and the affective polarity in a simile. These properties can be evoked from the vehicle, event, or the tenor component. While the vehicle is the most influential component, the roles of the event or tenor component are more limited, and often depend on their semantic richness.

For this research, I have created a simile data set consisting of 1,119 similes that are manually annotated with affective polarity. The similes are extracted from tweets with syntactic patterns that conform to the syntax of similes. They are then annotated using Amazon Mechanical Turk workers to obtain affective polarity judgements.

For this research, I have also created a simile data set consisting of 641 *open* similes. The similes are manually annotated with implicit properties. Using extraction patterns, open similes are first collected from tweets. They are then filtered by vehicles previously seen with explicit properties, so that the similes in the data set are more likely to have properties. Human provided implicit properties are then associated with the similes to create the gold

standard for implicit properties in similes. These data sets, annotated with affective polarity and implicit properties, are used in this research as the data sets for affective understanding of similes.

CHAPTER 5

RECOGNIZING AFFECTIVE POLARITY IN SIMILES

A simile describes the writer's subjective perception about a state or activity. This state or activity can be associated with a positive or negative affective polarity. But many similes do not contain words that have any obvious association with polarity. This makes it challenging to recognize the affective state evoked in a simile using existing sentiment resources, and presents the need for specialized methods. In this chapter, I present my research on automatically recognizing affective polarity in similes.

In Section 5.1, I will first present an overview of a supervised classification model for recognizing affective polarity of similes, the feature set used by the supervised classifiers, and implementation details. In Section 5.2, I will introduce baseline methods for classifying affective polarity using existing sentiment resources. In Section 5.3, I will show classification results with manually annotated training data. In Sections 5.4 and 5.5, I will present methods to automatically label training data for affective polarity classification, and classification results with the automatically labeled training data. Finally, in Section 5.6, I will present a qualitative analysis of the classifiers' behavior with respect to the figurative and literal nature of similes and will present some error cases.

5.1 Supervised Classification of Affective Polarity in Similes

5.1.1 Overview

The goal of this work is to recognize the general affective polarity of a simile regardless of any context in which it may appear (effectively, a *prior* polarity). Each simile is represented as a triple of the tenor, event, and vehicle component, since they are the three important building blocks of a simile (e.g., *"my face looks red like a tomato"* → <face, look, tomato>). The property of comparison can also be valuable information for determining the overall affective polarity of a simile. But it is an optional component, many similes[1] do not have it,

[1] 92% from previous estimate.

so it is not part of the instance representation used in this research. I present a supervised classification model that uses information that can be derived from different components of a simile. This information represent various lexical, semantic, and sentiment properties of the tenor, event, and vehicle components, given the triples as the simile instances. By looking at the simile triples in isolation, the goal is to determine their stereotypical affective polarity as perceived by most people in general, and not in any specific context.

The presented supervised classification requires training data containing simile triples that are labeled with positive, negative, or neutral polarity. Two binary classifiers are trained, one for each of the positive and negative polarity class. The classifiers use features that are extracted for each simile instance in the training data. Once the classifiers are trained, they are applied on the simile triples of the evaluation data to predict a positive or negative label for each simile. A separate classifier is not trained to predict the neutral category label. Instead, the neutral label is determined from postanalysis of the predictions by the positive and negative polarity classifiers (if neither predicts a polarity, or both predicts a polarity that is opposite of each other).

The training data for the supervised classifiers come from the data set described in Section 4.4. The similes in this data set are manually annotated with their affective polarity. In additional experiments I present in this work, the supervised classifiers also use training data that are automatically labeled. These data sets are created by exploiting any sentiment words present in a simile, or present in the surrounding contexts of many instances of a simile. The classification models trained using the manually annotated training data and automatically labeled training data are applied on evaluation data to judge how well they can classify affective polarity in new similes.

5.1.2 Feature Set for Supervised Classification

In this section, the feature set used by the supervised classifiers is described. The classifiers use three types of features from a simile, representing the lexical, semantic, and sentiment properties of the simile components.

5.1.2.1 Lexical Features

The lexical features look at the surface form of words and phrases in different components of a simile. These features are described below.

5.1.2.1.1 Unigrams. A binary feature indicates the presence of a unigram in a simile. This feature is not component specific, so the unigram can be from any simile component (i.e., tenor, event or vehicle). For example, three features are extracted from

the simile triple <bed, feel, rock>: "bed", "feel", and "rock". The feature values are binary, indicating presence or absence of these words in a simile.

5.1.2.1.2 Simile components. A binary feature is used for each tenor, event and vehicle phrase in the data set. This feature is component specific, i.e., the words are paired with their source component. For example, "dog" as a tenor is a different feature from "dog" as a vehicle in the similes <dog, be, friend> and <PERSON, obey, dog>, because they are extracted as *tenor:dog* from the first simile and *vehicle:dog* from the second simile. Also, this feature is not bag-of-words like the unigrams feature, and can take in multiword phrases when present in a simile component (e.g., *vehicle:ice box* in <room, feel, ice box>, *vehicle:bird's nest* in <hair, look, bird's nest>).

5.1.2.1.3 Paired components. A binary feature is used for each pair of simile components. The intuition here is that a pair of components may indicate affective polarity when used together. For example, *event:feel* and *vehicle:ice box* can be paired up as a feature because the pair can appear in multiple similes having negative polarity, but with different tenors such as house, room, or hotel. Similarly, the pair consisting of *tenor:person* and *vehicle:snail* can appear in similes with negative polarity, but with many different events such as move, run, or drive.

5.1.2.1.4 Explicit properties associated with vehicle. The explicit property is an optional component in a simile, and most similes do not have them. The instance representation used in this research (i.e., the triple <tenor, event, vehicle>) does not have the explicit property component, so that the classifiers can classify a simile even when an explicit property is not present in the simile. But given an instance that does not contain an explicit property, the classifiers can still look into the set of explicit properties that commonly appear with the simile's vehicle in the Twitter corpus to use as features.

To use the explicit properties as features, for each simile vehicle, all explicit properties mentioned with that vehicle across the Twitter corpus are extracted, and a binary feature represents each extracted explicit property. The intuition behind this feature is that some properties can be common across different similes, and may have the same effect on the affective polarity. For example, let's consider the following two similes: <Jane, swim, dolphin> and <Jim, run, cheetah>. If the property *fast* appears with both dolphin and cheetah as an explicit property in the Twitter corpus in similes where dolphin and cheetah are vehicles, then *fast* is extracted as a feature for the similes that contain these vehicles. Thus, if a simile with one of these vehicles appears in the training data, then the classifier can potentially learn to associate *fast* with positive polarity (in combination with other

features), and can apply the learned model on the new simile in evaluation data to predict its affective polarity.

5.1.2.1.5 Vehicle premodifiers. A binary feature is used for each noun or adjective pre-modifier that appears with the vehicle in similes in the Twitter corpus. The intuition is that the same pre-modifiers appearing with different vehicles can indicate the same affective polarity For example, knowing that <school, smell, *wet* dog> has a negative polarity in training data may help to classify a new simile, such as <elevator, smell, *wet* clothes>, as negative.

5.1.2.2 Semantic Features

The semantic features generalize the words of a simile to higher level semantic categories. The two types of semantic features used in this research are:

5.1.2.2.1 Hypernym class. Up to two levels of hypernym classes are used for each simile component head, using WordNet (Miller, 1995). For a word with multiple senses, only the first synset of the word is used. Once the hypernym classes are obtained for a word, the specific level information is no longer kept, and a binary feature represents each hypernym class. The intuition behind this feature is that groups of similar words can be used in different similes with the same affective polarity. For example, "room" is a hypernym of "bedroom" in WordNet, and they both have the hypernym "area". Knowing that <room, feel, Antarctica> is negative may help to classify <bedroom, feel, Antarctica> as also negative.

5.1.2.2.2 Perception verb. A binary feature indicates if the event component is a perception verb. Perception verbs are fairly common in similes (e.g., <IT, look, model>, <IT, smell, garbage>) as 48% of the similes in the development data (described in Section 4.4) contained a perception verb as the event. A set of the five most common perception verbs in similes (look, feel, sound, smell, taste) is used to determine the presence or absence of this feature. Although the feature is weak on its own, the intuition is that in combination with other features, it may be helpful.

5.1.2.3 Sentiment Features

The third and final type of features used are sentiment features. The sentiment features are designed to recognize words or phrases with positive or negative polarity in a simile, using existing sentiment resources. The MPQA Subjectivity Lexicon (Wilson et al., 2005) and the AFINN sentiment lexicon (Nielsen, 2011) were combined for this purpose.

5.1.2.3.1 Component Sentiment. Three binary features (one for each component) are used to indicate the presence of a positive sentiment word, and three binary features are used to indicate the presence of a negative sentiment word in each simile component. For example, <IT, feel, *heaven*> will have positive sentiment in the vehicle, but not in the tenor or event.

5.1.2.3.2 Explicit property sentiment. Two numeric features are used that count the number of positive and negative explicit properties that appear with the vehicle in the corpus. The property words are searched in the combined AFINN and MPQA lexicons to determine if they have any association with sentiment polarity. For example, if the vehicle "princess" from the simile <PERSON, look, princess> appears in the corpus with properties such as "pretty", "beautiful", "cute", "moody", etc., these properties are extracted as a set, and feature values are determined as three positive sentiment words and one negative sentiment word.

5.1.2.3.3 Predicted polarity by sentiment classifier for tweets. Two binary features (one for positive and one for negative) represent the label that a state-of-the-art sentiment classifier assigns to a simile. For this, I re-implemented the NRC Canada sentiment classifier (Zhu et al., 2014) using the same set of features described by the authors. For classification, a Java implementation[2] of SVM from LIBLINEAR (Fan et al., 2008) was used with the same parameter values used by the NRC Canada system. The sentiment classifier was trained with all of the tweet training data from SemEval 2013 Task 2 Subtask B (Nakov et al., 2013).

The sentiment classifier is designed to predict sentiment in tweets. To enable the classifier to classify polarity in similes, a simile is reconstructed from a triple by adding the comparator "like" (e.g., <face, look, tomato> → *"face look like tomato"*), but the words still remain lowercased and in their lemma form. The reason for reconstructing a simile from a triple is that the sentiment classifier uses N-gram features with N-gram lengths ranging from one to four. These two features account for if a state-of-the-art sentiment analysis system for Twitter data can recognize positive or negative polarity in a simile.

5.1.2.3.4 Simile connotation polarity. Two binary features (one for positive and one for negative) indicate the overall connotation of a simile. For this, the number of positive and negative connotation words in a simile are counted. If a simile contains more words with positive connotation than negative connotation, the overall connotation of the simile is determined as positive. If the simile contains more words with negative connotation than

[2] `http://liblinear.bwaldvogel.de/`

positive, the overall connotation of the simile is determined as negative. For identifying words with positive or negative connotation, a word connotation lexicon (Feng et al., 2013) is used.

5.1.3 Classification Model

As the supervised classification algorithm, a linear SVM classifier from LIBLINEAR (Fan et al., 2008) is used with its default parameter settings. The goal is to assign one of three labels to a simile: *Positive*, *Negative*, or *Neutral*. For this purpose, two binary classifiers are trained, one for positive and one for negative polarity. For the positive polarity classifier, similes labeled *positive* are used as the positive training instances. Similes labeled *negative* or *neutral* are used as the negative training instances. For the negative polarity classifier, similes labeled *negative* are used as the positive training instances, and similes labeled *positive* or *neutral* are used as the negative instances.

To classify a simile, both classifiers are used. If the simile is labeled as *positive* or *negative* (but not both), then it is assigned that label. If the simile is labeled as both *positive* and *negative*, or not labeled as either, then it is assigned a *neutral* label. A classifier is not designed to solely identify neutral similes because neutral similes are much less common than positive/negative similes, making up only 8.7% of the extracted similes in the development set (Table 4.9). Consequently, obtaining a large set of neutral similes via manual annotation would have required substantially more manual annotation effort.

5.2 Baseline Methods for Determining Affective Polarity in Similes

If similes commonly have one or more positive or negative sentiment words, sentiment lexicons can be used to recognize these words. Aggregated statistics about how many positive or negative sentiment words appear in a simile can be used to determine the overall affective polarity of the simile. It is also possible that a classifier that is designed with various features to recognize sentiment in text may be able to recognize affective polarity in a simile. But one of the motivations for this research is that many similes do not contain words having positive or negative polarity, and therefore existing sentiment resources are not sufficient to be used directly for recognizing affective polarity in similes. So for comparing how well the supervised classification model can classify similes with respect to their affective polarity, I present the following baseline methods that use various existing sentiment resources to determine affective polarity in similes.

5.2.1 Affective Polarity Determined Using the AFINN Sentiment Lexicon

The AFINN sentiment lexicon (Nielsen, 2011) contains 2,477 manually labeled words with integer values ranging from -5 (negativity) to 5 (positivity). For each simile, the sentiment scores for all lexicon words in the simile components are summed, and positive/negative polarity is assigned depending on whether the sum is positive/negative.

For example, for the simile <PERSON, talk, liar>, the vehicle "liar" has the sentiment score -3 in the lexicon, while the other component words do not appear. So the overall sum for the simile is -3, indicating that the simile has a negative polarity. If the overall sum is zero, or if the simile does not have any positive/negative sentiment word, the simile is classified as neutral.

5.2.2 Affective Polarity Determined Using MPQA Subjectivity Lexicon

The MPQA Subjectivity Lexicon (Wilson et al., 2005) contains 2,718 words with positive polarity and 4,910 words with negative polarity. The words are accompanied by their parts-of-speech. For a simile, first the part-of-speech for each word in the simile is determined using the CMU part-of-speech tagger for tweets (Owoputi et al., 2013). Then the words in the simile are searched in the MPQA lexicon after matching the parts-of-speech to determine how many positive sentiment or negative sentiment words appear in the simile. The overall positive/negative polarity of the similes is determined by if a simile has more positive/negative lexicon words.

For example, for the simile <PERSON, deceive, politician>, the event "deceive" is associated with negative polarity in the MPQA lexicon as a verb, and the other words do not appear. So there is a total of one word with negative polarity in this simile and no words with positive polarity. So the overall affective polarity of the simile is determined to be negative. A neutral label is assigned if the simile does not have any positive/negative sentiment words or has the same number of positive and negative sentiment words.

5.2.3 Affective Polarity Determined Using Connotation Lexicon

The connotation lexicon (Feng et al., 2013) contains 30,881 words with positive connotation and 33,724 words with negative connotation. To determine the overall affective polarity of a simile, the number of positive (and negative) connotation words in a simile are counted to determine if the number of positive connotation words are greater than the number of negative connotation words or the opposite. If a simile has more words with positive connotation, the simile is assigned a positive label. If the simile has more words

with negative connotation, the simile is assigned a negative label. As before, the neutral label is assigned if the simile has the same number of positive and negative connotation words or does not have any positive/negative connotation word.

5.2.4 Affective Polarity Determined Using Tweet Sentiment Classifier

The re-implemented NRC Canada sentiment classifier described in Section 5.1.2 is used to determine if a state-of-the-art tweet sentiment classifier can predict the affective polarity of a simile. After transforming a simile triple into a simile by adding the comparator "like", the sentiment classifier is applied to each simile in the data set. A positive/negative/neutral label is assigned to the simile depending on the predicted sentiment label by the classifier.

5.3 Classification Performance with Manually Annotated Data

Table 5.1 presents the results for supervised classification with the manually annotated data set (data set creation described in Section 4.4.2) using 10-fold cross-validation. The top section shows how effective these four existing sentiment resources are at assigning polarity to similes. Precision was sometimes very high. This suggests that when there are sentiment words in a simile that are easily recognizable, the overall polarity of the simile can be determined reliably. But recall for positive and negative polarity classes was low across the board, suggesting that many similes do not have sentiment words, and these similes cannot be easily identified using existing sentiment resources.

The lower section of Table 5.1 shows the results for the supervised classifiers. First, results for classifiers trained using only the sentiment features are presented in Row (a) in order to shed light on the effectiveness of traditional sentiment indicators as features with supervised learning. Row (a) shows that these classifiers produces reasonable precision (65-72%) but with recall levels only around 50% for both positive and negative polarity. But even with just the sentiment features, the recall scores are still much higher than any of the individual baseline methods. Recall is +16% higher for the positive polarity class, and +8% higher for the negative polarity class compared to the baseline method that yielded the best recall (Connotation Lexicon baseline).

Row (b) in Table 5.1 shows the results for a baseline classifier trained only with unigram features. Unigrams perform substantially better than the sentiment features for negative polarity, primarily due to a +22% recall gain, but only slightly better for positive polarity. Row (c) shows that the additional lexical features described in Section 5.1.2 further improve performance.

Table 5.1: Results with manually annotated training data (P = Precision, R = Recall, F = F1-score).

	Positive			**Negative**			**Neutral**		
	P	R	F	P	R	F	P	R	F
Sentiment Resource Baselines									
AFINN Lexicon	**88**	17	28	**95**	18	31	13	**95**	23
MPQA Lexicon	83	21	34	90	15	26	13	**95**	24
Connotation Lexicon	61	38	47	63	40	49	17	63	26
Re-implemented NRC Canada Sentiment Classifier	72	34	47	94	16	27	13	83	23
Affective Polarity Simile Classifiers									
(a) Sentiment Features	65	54	59	72	48	58	19	37	25
(b) Unigrams	73	52	61	74	70	72	21	47	29
(c) Unigrams + Other Lexical	73	56	63	75	76	75	**26**	45	**33**
(d) Unigrams + Other Lexical + Semantic	68	59	63	76	72	74	24	40	30
(e) Unigrams + Other Lexical + Semantic + Sentiment	75	**60**	**67**	77	**79**	**78**	25	40	31

Row (d) shows that adding the semantic features did not improve performance. One reason could be that some WordNet hypernym classes are very specific and may not generalize well. For example, <PERSON, run, snail> and <PERSON, run, turtle> have the same polarity because they both indicate running slowly. But the immediate hypernym classes for snail are *gastropod* and *mollusk* in WordNet, whereas for turtle, they are *chelonian* and *anapsid*. These hypernym classes are too specific to be common between these two vehicle words and will not be helpful.

Finally, Row (e) shows that adding the sentiment features along with all the other features yields a precision gain for positive polarity and a recall gain for negative polarity. Overall, the full feature set improves the F score from 61% to 67% for positive polarity, and from 72% to 78% for negative polarity, over unigrams alone.

For the neutral category, precision is low with the baseline methods (ranging from 13% to 17%), and improves slightly with the supervised classification (ranging from 19% to 26%), but still remains low. This shows that a good number of similes with positive or negative polarity are not recognized by the methods, and they are classified as neutral similes instead.

Table 5.2 presents a sample of similes where the vehicle terms appear only once in the data set. The tenor and the event words in these similes are commonly used with both positive/negative categories and mostly do not have any obvious association with positive or negative polarity. The unigram-based classifier could not classify these instances, but the classifier with the full feature set could. So it had to generalize beyond the surface-level lexical form of the component words by relying on some of the additional features other than just the unigrams.

In order to understand how much the size of the training set matters, the learning

Table 5.2: Similes with unique vehicle terms that were correctly classified using the full feature set.

Positive	Negative
<PERSON, feel, superhero>	<PERSON, feel, old woman>
<PERSON, be, friend>	<PERSON, be, hurricane>
<beast, look, beauty>	<IT, feel, eternity>
<PERSON, feel, hero>	<PERSON, feel, peasant>
<PERSON, feel, champion>	<PERSON, eat, savage>
<PERSON, seem, sweetheart>	<PERSON, be, witch>
<IT, be, sleepover>	<PERSON, feel, prisoner>
<IT, be, reunion>	<IT, be, north pole>
<PERSON, feel, president>	<IT, feel, winter>
<ronaldo, be, messi>	<PERSON, be, wolf>

curves of the classifiers using varying amounts of manually annotated data are presented in Figure 5.1. The results are shown for the classifiers trained only with unigram features and classifiers trained with the full feature set. The results are produced from 2-fold, 3-fold, 5-fold, and 10-fold cross-validation experiments, with the size of the corresponding training sets shown on the X-axis.

The trends in Figure 5.1 show that the classifiers with unigram features hit a plateau at about 600 training instances for both positive and negative classes. However, the classifiers with the full feature set continually benefited from more training data.

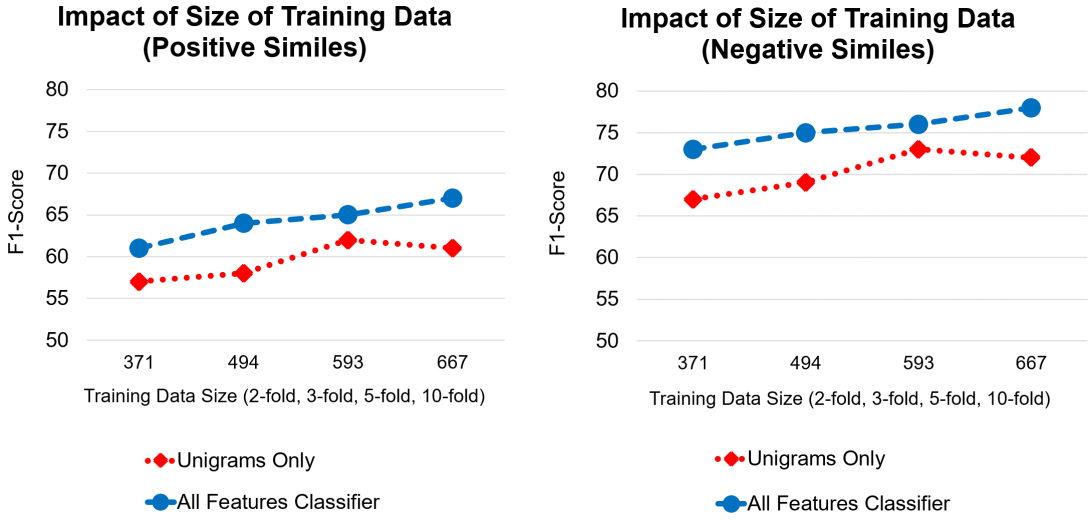

Figure 5.1: Learning curve for positive and negative similes.

5.4 Automatically Acquiring Labeled Training Data

Obtaining training data with manual annotation is time consuming and may not always be readily available for a different domain. If affective polarity of some types of similes is easy to recognize and these similes can be automatically labeled with their affective polarity, they can be given to a supervised classifier to use as training instances. Therefore, in this research, I also present methods to obtain labeled training data automatically for affective polarity classification of similes.

In Section 5.2, I introduced baseline methods that use existing sentiment resources to classify similes into affective polarity classes. As similes sometimes have words with strong polarity (e.g., <bed, feel, *heaven*> or <PERSON, *cry*, baby>), these words can potentially be identified using existing sentiment resources. The baseline results in Section 5.3 demonstrated that although these methods are not always able to recognize similes with affective polarity, when they do, they do it with good precision (except for connotation lexicon). The automatic methods I introduce in this section use similar techniques, but on a large collection of unannotated similes. The hope is that the automatic methods will be able to identify the polarity of many similes from the unlabeled data so that a large training data can be easily created without manual supervision. Because sentiment resources have limitations (e.g., sentiment classifiers are not perfect, sentiment lexicons do not possess knowledge of context), these training instances can be expected to contain some noise. But for any new domain (e.g., Amazon product reviews), being able to automatically obtain training instances can be valuable.

I created six types of additional training data sets that are automatically labeled with affective polarity, so that it can be judged how well the supervised classification works with automatically labeled data compared to classification with training data that are manually annotated. These data sets were created from 7,594 similes (Section 4.4.1) appearing in ≥ 5 tweets in the Twitter corpus (excluding the 1,500 manually annotated similes). They were created using existing sentiment resources, with the following methods.

5.4.1 Using AFINN Sentiment Lexicon Words

The first training data set was created using the AFINN sentiment lexicon (Nielsen, 2011) which contains 2,477 manually labeled words with integer values ranging from -5 (negativity) to 5 (positivity). For each simile, the sentiment scores for all lexicon words in the simile components were summed, and positive/negative polarity was assigned depending on whether the sum is positive/negative. This method yielded 460 positive and 423 negative similes.

5.4.2 Using MPQA Sentiment Lexicon Words

The second training data set was created using the MPQA Subjectivity Lexicon (Wilson et al., 2005) that contains 2,718 positive words and 4,910 negative words. After matching parts-of-speech for the words against the MPQA lexicon, the positive/negative polarity of the similes were determined by if a simile had more positive/negative lexicon words. This method yielded 629 positive and 522 negative similes.

5.4.3 Using Sentiment Classifiers

The third training data set was created using the re-implemented NRC Canada sentiment classifier (Zhu et al., 2014) which uses the same set of features as well as the same parameter values used in their original work. A simile is labeled as positive or negative if the sentiment classifier labeled it as positive or negative, respectively. This method yielded 1,185 positive and 402 negative similes.

5.4.4 Using Sentiment in Surrounding Words

The previous approaches for labeling training instances primarily identified similes that contained one or more strongly affective words. This can potentially bias the training data and limit the classifier's ability to learn to recognize affective similes that do not contain words with a positive or negative connotation. Therefore, I explored an additional approach where instead of judging the sentiment of the words in a simile, the words in the tweet surrounding the simile are analyzed.

Intuitively, there are often redundant sentiment indicators in a tweet. For example, *"I **hate** it when my room is as cold as Antarctica"* already contains a negative sentiment word "hate", which indicates that the state of the described room is negative. For each simile in the data set, all tweets that contained the simile were identified and all of the words surrounding the simile in these tweets were combined as a collective "context" for the simile. One issue is that when people feel amused (e.g., *"he looks like a complete zombie, haha"*) or sarcastic (e.g., *"my room feels like an igloo. great! LOL."*), seemingly positive words in the context can be misleading because the sentiment is actually negative. As a simple measure to mitigate this issue, a small set of laughter indicators ("lol", "haha", and up to four repeated occurrences of "ha") were manually removed from the lexicons.

For each simile, the number of distinct positive and negative sentiment words in the collective context were then counted, and the ratio of positive (or negative) sentiment words to all the sentiment words in the collective context were calculated as a score for the simile. The simile was then labeled with the corresponding polarity if the score was

higher than a threshold (here, 0.7 to ensure high quality). As the sentiment lexicon, words from the MPQA subjectivity lexicon and AFINN sentiment lexicon were combined. By analyzing sentiment in the surrounding contexts of a simile, the hope is that this method would bring in a more diverse set of training instances than the methods that would only identify positive/negative sentiment words in the simile components. This method yielded 492 positive and 181 negative similes.

5.4.5 Combination of Training Instances

I created two additional training sets by combining sets of instances labeled using the different methods above. As the fifth set, I combined training instances collected using the MPQA and AFINN lexicons and the NRC Canada sentiment classifier, which yielded a total of 1,429 positive similes and 754 negative similes. As the sixth set, I added the instances recognized from the surrounding words of a simile to the fifth set, producing the largest data set of 1,724 positive and 874 negative similes.

The supervised classifiers also need neutral similes to use as negative instances. But there is no straightforward or trivial way to reliably identify neutral similes automatically. The automatic methods for acquiring training instances described in Section 5.4 were designed mainly for identifying positive and negative training instances.

Therefore, additional presumed neutral instances were added by randomly selecting instances that were not identified as either positive or negative by the automatic methods and also did not contain a sentiment lexicon word in their surrounding words. However, some noise can be expected because not being recognized as a positive or negative simile using the automatic methods does not guarantee that a simile will be neutral.

Also, the positive, negative and neutral instances acquired using the automatic methods varied substantially from the class distribution in the development data. Therefore, instances were randomly selected from the classes in the final training sets, maintaining the class size ratio of the development data. The final training data sizes for each automatic approach is reported in Table 5.3.

5.5 Classification Performance with Automatically Acquired Training Data

Table 5.4 shows the performance of the classifiers (using the full feature set) when they are trained with automatically acquired training instances. The upper section of Table 5.4 shows results using training instances labeled by three different sentiment resources. Rows (a) to (c) show that precision ranges from 65% to 78% for the positive polarity and 81% to

Table 5.3: Final training set sizes for automatically labeled data.

Methods	# of Training Instances		
	Positive	**Negative**	**Neutral**
(a) labeled data using AFINN	384	423	78
(b) labeled data using MPQA	475	522	94
(c) labeled data using NRC Canada	365	402	74
(d) labeled data from (a), (b), + (c)	686	754	136
(e) labeled data using sentiment in surrounding words	164	181	34
(f) labeled data from (a), (b), (c), + (e)	795	874	158

Table 5.4: Results with automatically labeled training data (P = Precision, R = Recall, F = F1-score).

Classifier	**Positive**			**Negative**			**Neutral**		
	P	R	F	P	R	F	P	R	F
(a) SVM with labeled data using AFINN	**78**	32	45	85	31	45	**14**	80	**24**
(b) SVM with labeled data using MPQA	65	44	53	81	27	41	12	59	20
(c) SVM with labeled data using NRC Canada	72	34	47	**94**	16	27	13	**83**	23
(d) SVM with labeled data from (a), (b), + (c)	65	56	60	86	30	45	14	58	22
(e) SVM with labeled data using sentiment in surrounding words	60	57	59	62	**57**	**60**	13	20	16
(f) SVM with labeled data from (d) + (e)	61	**61**	**61**	70	52	**60**	12	24	16

94% for the negative polarity. Recall ranges from 32% to 44% for the positive polarity and 16% to 31% for the negative polarity. Row (d) shows that combining the training instances labeled by all three resources produces the best results for the positive polarity class by improving the F1-score by +7% over the classification that uses training instances acquired using the MPQA lexicon (Row (b)). The results are similar (both precision and recall) for the negative class when compared with the classification that uses training instances acquired using the AFINN lexicon (Row (a)).

Row (e) of Table 5.4 shows the performance of the classifiers when they are trained with instances selected by analyzing sentiment in the surrounding words of the similes. This results in a substantial recall gain over any other individual methods (Row (a), (b) and (c)), which validates the hypothesis that similes obtained by recognizing sentiment in their surrounding words provide the classifier with a more diverse set of training examples. Finally, Row (f) shows that using both types of training instances further improves performance for positive polarity, and increases precision for negative polarity but with some loss of recall.

Table 5.5 summarizes the results for comparison. The upper section of the table shows results for the baseline methods from Table 5.1. The lower section presents results for the affective polarity classifiers with manually and automatically labeled data using the full feature set. These results show that there is still a gap between the performance of the classifiers trained with manually annotated data versus automatically acquired data. However, the classifiers trained with automatically acquired data produce substantially higher F1-scores than all of the baseline systems in Table 5.1. The use of automatically acquired training data is a practical approach for creating simile classifiers for specific domains, such as Amazon product reviews (e.g., *"headphone sounds like garbage"*, or *"each song is like a snow-flake"*) studied in previous work on figurative comparisons in similes (Niculae and Danescu-Niculescu-Mizil, 2014).

5.6 Analysis and Discussion

In this section, I present a qualitative analysis of the simile corpus and the behavior of the classifiers. There can be at least two reasons why similes might be difficult to classify with respect to their affective polarity. First, the interpretation of a simile can be highly context-dependent and subjective, depending on the speaker or the perceiver. To illustrate, Table 5.6 presents examples of similes that can have different polarity depending on the speaker or perceiver's personal experience or location, and other subjective aspects of the context.

The similes are accompanied by imaginary context that would support the corresponding polarity assignments. For example, <IT, look, snow> may be a good thing to someone who lives in Utah where people look forward to skiing, but a bad thing to someone living in Boston during the unusually snowy winter of 2015. <IT, look, rain> can be viewed positively by people who live in drought-stricken California, but perhaps not by people who

Table 5.5: Comparison of results (P = Precision, R = Recall, F = F1-score).

Classifier	**Positive**			**Negative**			**Neutral**		
	P	R	F	P	R	F	P	R	F
Sentiment Resource Baselines									
AFINN Lexicon	**88**	17	28	**95**	18	31	13	**95**	23
MPQA Lexicon	83	21	34	90	15	26	13	**95**	24
Connotation Lexicon	61	38	47	63	40	49	17	63	26
Re-implemented NRC Canada Sentiment Classifier	72	34	47	94	16	27	13	83	23
Affective Polarity Classifiers									
SVM with automatically labeled data	61	**61**	61	70	52	60	12	24	16
SVM with manually labeled data	75	60	**67**	77	**79**	**78**	**25**	40	**31**

Table 5.6: Example of similes that can potentially have different polarity in different context.

Simile	Imagined Context	Polarity
<PERSON, smell, baby>	young mother	positive
<PERSON, smell, baby>	MTurkers	negative
<IT, look, snow>	lives in Boston	negative
<IT, look, snow>	lives in Utah	positive
<IT, look, rain>	lives in England	negative
<IT, look, rain>	lives in California	positive

live in perpetually rainy England. <PERSON, smell, baby> can be positive to new mothers, but was viewed as negative by the Mechanical Turk annotators.

Second, the polarity of a simile may interact with the distinction made in previous work between figurative and literal uses of similes (Bredin, 1998; Addison, 1993). For example, Niculae and Danescu-Niculescu-Mizil (2014) showed that sentiment and figurative comparisons are strongly correlated. Thus, most literal comparisons are neutral while most figurative comparisons carry polarity. To explore this issue, in collaboration with Professor Marilyn Walker, University of California Santa Cruz, an informal analysis of the 378 similes in the development data set was conducted to examine the literal vs. figurative distinction. For this analysis, both the simile component triples as well as the context of ten tweets in which the simile appeared were considered.

The findings suggest that, 1) the distinction between positive and negative similes in the data is orthogonal to the figurative vs. literal distinction; 2) some similes are used both figuratively and literally, and cannot be differentiated without context; 3) even in cases when all samples were literal, it is not challenging to invent contexts where the simile might be used figuratively, and vice versa; and 4) for a particular instance (simile + context), it is usually possible to tell whether a figurative or literal use is intended by examining the simile context (although some cases remain ambiguous). Table 5.7 shows examples of some similes that are identified in the analysis as being figurative, literal, or both depending on context. For example, the simile <PERSON, look, frankenstein> appeared in figurative context referring to someone who did not have makeup that day, and also in literal context referring to an actor playing the role of Frankenstein in theater.

These observations reinforce the difficulty with making the figurative/literal distinction noted by Niculae and Danescu-Niculescu-Mizil (2014), whose annotation task required Turkers to label comparisons on a scale of 1 to 4 ranging from very literal to very figurative. Even with Master Turkers, a qualification task, filtering annotators by gold standard items,

Table 5.7: Similes with figurative or literal interpretation, or ambiguous depending on the context.

Use	Polarity	Simile
figurative	positive	<house, smell, heaven>
figurative	positive	<PERSON, look, queen>
figurative	negative	<PERSON, look, tomato>
literal	negative	<hair, smell, smoke>
literal	neutral	<PERSON, look, each other>
both	neutral	<house, smell, pizza>
both	negative	<PERSON, look, skunk>
both	negative	<PERSON, look, frankenstein>

and collapsing scalar 1,2 values to literal and 3,4 values to figurative, the interannotator agreement with Fleiss' κ was 0.54. They note that out of 2400 automatically extracted comparison candidates, only 12% ended up being selected confidently as figurative comparisons.

Selected cases that the supervised classifiers fail on are further illustrated in Table 5.8. Examples S1 to S3 could be related to the difficulties noted above with subjectivity of interpretation. Many people for example like the smell of coffee and pizza, but perhaps not when a person smells that way. In examples S4 to S7, the subjective interpretation can come from an inferred property of a simile. For example, babies are often attributed with positive characteristics such as cute, adorable, innocent, etc. In similes with events such as sleep or feel, these are the properties that would apply in most circumstances, and the similes are typically viewed to have positive polarity. But in similes with events such as smell and sound, the polarity can be negative because of a different set of inferred properties that are specific to other circumstances (e.g., the baby needs a diaper change, or cries loudly). The

Table 5.8: Error analysis of classifier output (Man = classifier trained with manually annotated instances, Auto = classifier trained with automatically annotated instances).

ID	Simile	Gold	Man	Auto
S1	<PERSON, smell, coffee>	negative	positive	positive
S2	<PERSON, smell, pizza>	positive	negative	neutral
S3	<IT, smell, pizza>	neutral	positive	neutral
S4	<PERSON, sleep, baby>	positive	positive	neutral
S5	<PERSON, smell, baby>	negative	negative	neutral
S6	<PERSON, feel, baby>	positive	negative	positive
S7	<PERSON, sound, baby>	negative	positive	neutral
S8	<PERSON, sound, pirate>	positive	negative	negative
S9	<PERSON, look, pirate>	negative	negative	neutral

positive or negative interpretation of *sounding like a pirate* and *looking like a pirate* may also be context dependent in examples S8 and S9 (e.g., looking like a pirate may be cool on Halloween but not otherwise).

5.7 Chapter Summary

In this chapter, I presented a supervised classification model for recognizing affective polarity in similes. The features for the supervised classifiers are derived from the three major simile components: tenor, event, and vehicle. The features represent lexical, semantic, and sentiment properties of these components. The classifiers assign a positive, negative, or neutral label to similes.

Existing sentiment resources are used to assign a positive, negative, or neutral label to a simile as baseline methods. As the existing sentiment resources, sentiment lexicons, a connotation lexicon, and a sentiment classifier designed for tweets are used. The lexicon based approaches identify positive/negative words in simile components and aggregate statistics for each simile to assign a label. The sentiment classifier use features applicable for tweet sentiment classification for predicting the affective polarity of a simile. The results show that existing sentiment resources are insufficient and cannot be easily used to recognize affective polarity for most similes.

Supervised classifiers are trained with manually annotated data for affective polarity classification. The manually annotated data set is relatively small but of high quality. The results demonstrate that substantial improvements can be achieved in classification using the proposed feature set over the baseline methods that classify similes using existing sentiment resources.

Existing sentiment resources are used to identify positive/negative sentiment words or the overall sentiment polarity in a simile to automatically label training instances for supervised classification. They are similar to the baseline methods for affective polarity classification, but they are applied to unlabeled similes to automatically acquire training data for classification. These data sets are larger in size than manually annotated data set but contain some noise. Results demonstrate that good classification performance can also be achieved with automatically labeled training data with the proposed feature set, which are substantially better than the baseline methods that use existing sentiment resources. However, compared to results with manually annotated training data, results with automatically labeled training data are not as good, leaving room for improvement in future work. But when manually labeled data may not be available, this approach represents an alternative option for training supervised classifiers.

Analysis of the data suggests that figurative versus literal distinction in similes is orthogonal to the positive versus negative polarity in similes. Also, figurative versus literal use of similes depend on the context. A simile can appear in context suggesting a figurative comparison, but the same simile can also appear in context that would suggest a literal comparison. The polarity of a simile may depend on the writer's or reader's personal experience or location, and other subjective aspects of the context.

CHAPTER 6

INFERRING IMPLICIT PROPERTIES IN SIMILES

An explicit or implicit property is an important component in a simile. It works as the basis of a comparison, but it is an optional component. Many similes, which are known as the *open* similes, do not explicitly mention the property of comparison. They form the vast majority of the similes (92%) in tweets, and for them, the property must be inferred. Being able to infer the implicit property is fundamental for affective understanding of similes, and can have downstream contributions for affective polarity recognition. In this chapter, I present a framework for automatically inferring implicit properties in similes.

In Section 6.1, I first introduce the framework for inferring implicit properties in similes. In Section 6.2, I present a variety of methods to generate initial candidate properties and present statistics about the productivity and coverage of these methods. In Section 6.3, I present methods to rerank the initial candidate properties using influence from complementary components. In Section 6.4, I present aggregate ranking of the individual methods and the final results for implicit property ranking. In Section 6.5, I present an analysis on the difficulties associated with similes with different levels of interpretive diversity (i.e., how many different properties a simile has). Finally in Section 6.6, I discuss additional experiments for using the inferred properties to improve affective polarity classification results.

6.1 Overview of the Property Inference Framework

For inferring properties in *open* similes, I decompose the problem into three subtasks: (1) generating candidate properties, (2) evaluating the candidate properties with respect to multiple simile components, and (3) aggregated ranking of the properties. Figure 6.1 illustrates the approach.

First, the vehicle and event components of a simile are used individually to generate candidate properties. For this purpose, I investigate a variety of candidate generation methods, including harvesting properties from syntactic structures and dictionary defini-

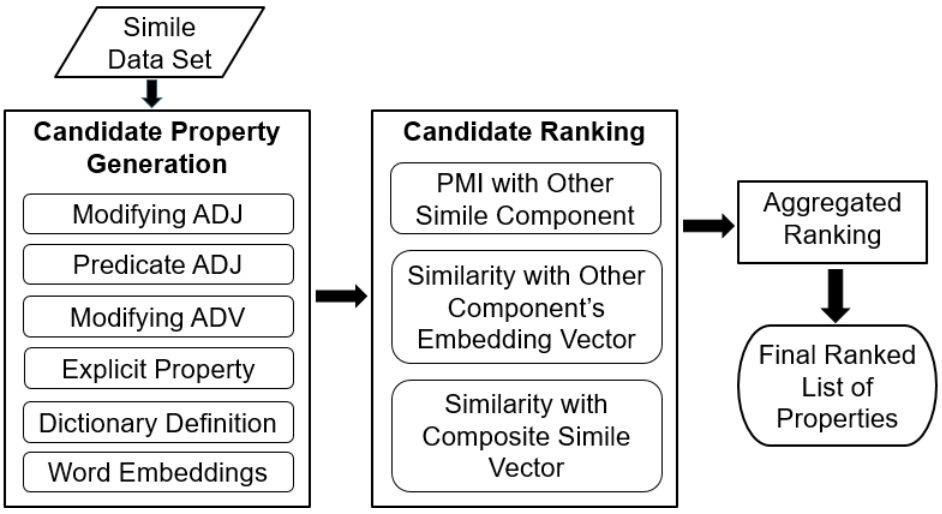

Figure 6.1: Framework for inferring implicit properties.

tions, identifying relevant properties using statistical co-occurrence, and assessing similarity between word embedding vectors.

Second, the candidates generated by each method are evaluated based on their strength of association with the complementary component of the simile (event or vehicle), to assess their compatibility with the complementary component. For candidates generated from the vehicle term, the candidates are evaluated based on their association with the event term, and vice versa. Three association measures are explored for this purpose: point-wise mutual information (PMI) to measure statistical co-occurrence, and vector similarity using single and composite word embeddings.

Third, an aggregated ranking is produced over the entire set of properties hypothesized by *all* of the candidate generation methods. Intuitively, each candidate generation method is viewed as an independent source, and the collective evidence looks across the set of different candidate generation methods (similar to an ensemble). Each property is scored based on its average rank across the different methods, so that properties highly ranked by multiple methods are preferred. In the following section, I describe each step of this process in more detail.

6.2 Candidate Property Generation

Candidate properties are generated from the vehicle and event words of a simile.[1] However, when the event is a form of "to be" or a perception verb (taste, smell, feel,

[1]Although the tenor can also sometimes generate candidates, this seemed to be relatively rare in the development data set, so I did not use the tenor in this research.

sound, look), candidate properties are not generated from the event because the verb is too general. Only 73 (16%) of the similes in the data set (described in Section 4.5) had a verb other than "to be" or a perception verb. The properties are restricted to be adjectives, adverbs, or verb forms that can function as nominal premodifiers (e.g., "crying baby", "wilted lettuce").

6.2.1 Methods

A total of seven methods are used for generating candidate properties. These methods are applied to generate candidate properties from the entire Twitter corpus of roughly 140 million tweets, and not just the simile data set.

6.2.1.1 Modifying ADJ

Given a vehicle term of a simile, premodifying adjectives of the vehicle term are extracted from the Twitter corpus. For example, the vehicle term is "tomato" in the simile <face, look, tomato>. For this example, one of the candidate properties "ripe" may be extracted for the vehicle "tomato" if the phrase "ripe tomato" appears in the Twitter corpus.

6.2.1.2 Predicate ADJ

Given a vehicle term, adjectives in predicate adjective constructions with the vehicle are extracted. For example, "red" is extracted for the vehicle "tomato" from the phrase "tomato is red".

6.2.1.3 Modifying ADV

Given an event term (verb), adverbs that precede or follow the verb are extracted. For example, "immaturely" is extracted for the event "act" due to the phrase "acts immaturely" or "immaturely acts" in the Twitter corpus.

6.2.1.4 Explicit Property

Properties that are mentioned explicitly in comparison phrases are also extracted as candidate properties. For vehicle terms, phrases of the form: "ADJ/ADV like $\text{NP}_{vehicle}$" (e.g., *"cold like Antarctica"*) and "ADJ/ADV as $\text{NP}_{vehicle}$" (e.g., *"cold as Antarctica"*) are identified, and candidate properties are extracted from the ADJ/ADV part of the phrase . For event terms, properties are extracted from phrases of the form: "VERB_{event} ADJ/ADV like" (e.g., *"feels cold like"*) and "VERB_{event} as ADJ/ADV as" (e.g., *"feels as cold as"*).

6.2.1.5 Dictionary Definition

Dictionary definitions often mention salient properties associated with a word. From the definitions of the vehicle and event terms, adjectives, adverbs, and verbs (functioning as premodifiers) are harvested as candidate properties. The definitions are acquired using Wordnik[2], which contains 5 source dictionaries: Heritage Dictionary of the English Language, Wiktionary, the Collaborative International Dictionary of English, The Century Dictionary and Cyclopedia, and WordNet 3.0 (Miller, 1995).

6.2.1.6 PMI

Given a vehicle or event term, point-wise mutual information (PMI) is computed between that term and adjectives, adverbs, and verbs functioning as an adjective (appearing in $\geq$ 100 tweets) in the Twitter corpus. Terms with high PMI scores are used as candidate properties.

6.2.1.7 Word Embeddings

The final method for generating candidate properties uses word embedding vectors. Word embedding vectors are low dimensional vector representations of words in a corpus. For this purpose, a word embedding model is trained using the Twitter corpus, limiting the vocabulary to nouns, verbs, adjectives, and adverbs that occurred in $\geq$ 100 tweets. For training, word2vecf[3] (Levy and Goldberg, 2014) is used which allows training for arbitrary contexts using the skip-gram model.[4] The model outputs two types of vectors: word vector and context vector. For each types of vectors, 300 dimensions are used.

Candidate properties are then generated by selecting the words whose context vector is most similar to the vehicle or event's word vector using cosine similarity. This is because properties are expected in the context of a component word (e.g., "cold" can be expected in the context of Antarctica when inferring a property for the simile <room, feel, Antarctica>). Similarity between a word vector and a context vector is comparable to first order similarity (Levy et al., 2015), whereas the more traditionally calculated similarity between two word vectors is comparable to second order similarity. To control for noisy candidates, a candidate

---

[2] `https://www.wordnik.com/`

[3] `https://bitbucket.org/yoavgo/word2vecf`

[4] In their method, context of a word can be defined selectively, such as by words co-occurring in dependency relations.

property needs to occur with the vehicle (or event) as a bigram (in any order) with frequency ≥ 10 in the Twitter corpus.

With each method, the candidates are ranked and the top 20 properties are selected. For the four methods that use syntactic patterns, P(property | source component) is calculated based on the number of times the property and the source component (i.e., vehicle or event) appear together in that syntactic construction relative to all times the source component appears in that syntactic construction. The probability is then used to rank the candidates. For the dictionary definition method, the properties are sorted based on how many of the 5 dictionaries mention the property in the word's definition. Ties are broken based on the frequency of the property in the definitions. For the word embedding-based methods, cosine similarity scores are used for ranking.

6.2.2 Productivity of the Candidate Generation Methods

With the candidate generation methods, first the number of candidates each method is able to generate needs to be analyzed. If a method generates too few candidates, it will not be very useful. Conversely, if a method generates a large number of candidates, then the ranking framework needs to be robust to rank the plausible properties higher than the properties that do not fit.

Table 6.1 presents average, minimum, and maximum number of candidates generated by the different candidate generation methods. The PMI and Word Embedding-based methods were excluded here as these methods generate a ranking of all words in the corpus as candidate properties. The methods that use the explicit property extraction patterns and dictionary definitions generated fewer candidates than the methods that use general

Table 6.1: Statistics about candidates generated by different methods. Similes with a "to be" or perception verb were excluded for the methods that use the event as the source.

	Average	Minimum	Maximum
# of Candidates Generated from Vehicle			
Modifying ADJ	423.62	1	3177
Predicate ADJ	104.21	0	1070
Explicit Property	8.28	0	116
Dictionary Def.	20.5	0	71
# of Candidates Generated from Event			
Modifying ADV	68.67	2	223
Explicit Property	19.85	0	61
Dictionary Def.*	18.59	3	55

syntactic structures. Average number of candidates generated by these methods ranged from 8.28 to 20.5, and maximum number of candidates by these methods ranged from 55 to 71. The method that uses modifying adjectives generated more candidates than any other method, averaging 423.62 candidate properties, and reaching up to a maximum of 3,177 candidate properties. All of the methods generated fewer than five candidate properties for some similes as shown by the "Minimum" column.

Figure 6.2 presents the percentage of similes that have at least K candidates generated by different candidate generation methods. As before, the PMI and Word Embedding-based methods were excluded. The trend lines in the figure show that the methods that use explicit property extraction patterns and dictionary definitions do not generate more than 20 candidate properties for most similes (for less than 50% in the data set). The method that uses modifying adjectives generated more than 100 candidate properties for about 80% of the similes. These statistics show that most of the candidate generation methods are productive as they were able to generate at least 10-20 candidates for many similes.

6.2.3 Coverage of the Generated Candidates

Next, it is important to explore the effectiveness of the candidate generation methods. One of the primary concerns here is to assess whether candidate generation methods are able to generate at least some acceptable properties. They can be expected to over-generate, but they need to produce at least one acceptable property or the downstream components will

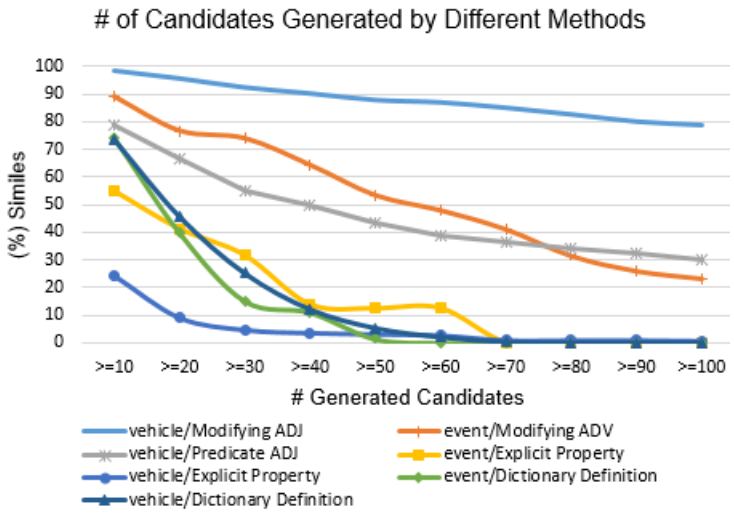

Figure 6.2: Percentage of similes that have at least K candidates generated by different methods. Similes with a "to be" or perception verb were excluded for the methods that use the event as the source.

be helpless. To assess this, the coverage of each candidate generation method is evaluated based on the Top 10, Top 20, and Top 30 properties that it produced. Coverage is the percentage of similes for which the method generates at least one gold standard property (from the human annotators).

Table 6.2 shows that the *Dictionary Definitions* for vehicles is the best performing method for the Top 10 candidates, generating at least one acceptable property for 40% of the similes. The *Modifying ADJ* method performs best for the Top 30 candidates, generating an acceptable property for 63% of similes. Note that the *Explicit Property* method performs reasonably well (40% coverage for Top 30 properties generated from vehicles and 6% coverage for properties generated from events), but clearly is not sufficient on its own, showing the limitation of harvesting explicitly stated properties.

The ALL rows show the coverage obtained by combining the property lists from all generation methods listed above in the table. The combined set of properties (Top 30) generated from vehicles yields 86% coverage, while the combined set of properties generated from events yields only 10% coverage (partly because these methods apply to only 16% of the similes), showing that vehicles are more effective for candidate generation.

Table 6.2: Coverage and MRR for the candidate generation methods. Top10, Top20, Top30 = percent of similes with a plausible property within top 10, 20, 30 ranked properties. Methods excluded in "ALL" and "TOTAL" rows are marked with (*). In the MRR calculation when the event component is the source, similes with a "to be" or a perception verb were excluded.

	Top10	Top20	Top30	MRR
Coverage of Candidates Generated from Vehicle				
PMI*	18%	31%	37%	.06
Modifying ADJ	39%	55%	63%	.16
Predicate ADJ	28%	39%	43%	.11
Explicit Property	37%	39%	40%	.23
Dictionary Def.	40%	47%	49%	.22
Word Embedding	35%	48%	58%	.15
ALL	76%	84%	86%	n/a
Coverage of Candidates Generated from Event				
PMI*	2%	3%	4%	.09
Modifying ADV	4%	5%	5%	.13
Explicit Property	4%	5%	6%	.16
Dictionary Def.*	3%	4%	4%	.09
Word Embedding	5%	6%	6%	.16
ALL	9%	10%	10%	n/a
All Candidates				
TOTAL	78%	86%	88%	n/a

However, the TOTAL row shows that combining properties generated from both vehicles and events yields 88% coverage using the Top 30 candidates. The Top 20 candidates provide coverage that is nearly as good (86%) with substantially fewer properties to process downstream. So the Top 20 candidates are used for the later steps of the experiments.[5]

The last column of Table 6.2 shows candidate ranking results based on Mean Reciprocal Rank (MRR) for the top 20 properties produced by each candidate generation method and ranked by their initial ranking criteria. MRR is calculated by (here, S is the set of similes):

$$MRR = \frac{1}{|S|} \sum_{s \in S} \frac{1}{(rank\ of\ 1^{st}\ acceptable\ property\ for\ s)}$$

The coverage statistics in Table 6.2 demonstrated that the candidate generation methods are largely sufficient to find at least one acceptable property for most similes. However, they also generate many unacceptable properties. The MRR results showed that when using the initial ranking criteria of the candidate generation methods, acceptable properties are not ranked at the top positions most of the times. So the next challenge is to identify which of the candidate properties are more appropriate for a given simile than the rest of the candidate properties. The *PMI* method (for both vehicles and events) and the *Dictionary Definition* method (for events) produced low MRR scores < 0.10. Therefore, these candidate generation methods were not used in later steps.[6]

6.3 Reranking the Candidate Properties Using Influence from the Second Component

Next, the goal is to investigate whether the initial ranking results in the previous step can be improved by considering the second component of the simile. Intuitively, suppose that "green", "slow", and "endangered" are generated as candidate properties from the vehicle "turtle" (e.g., for <dad, drive, turtle>. Taking the event verb "drive" into account can help to rank "slow" more highly than the other candidates. In this section, I present methods to rerank the candidate properties.

---

[5]The decision to use the Top 20 candidates was based on similar results on the development data.

[6]This decision was based on similar results observed on the development data.

6.3.1 Methods for Reranking

Three criteria are used to rank candidates generated from a simile component based on its association with the second component (unless the event is "to be", in which case the original candidate ranking is retained because the verb is too general).

6.3.1.1 Pointwise Mutual Information (PMI)

As the first ranking criteria, Pointwise Mutual Information between a candidate property and the second component of a simile is calculated. This method is referred to as: **PMI**.

6.3.1.2 Word embedding vector similarity

The trained word embedding model is used to calculate cosine similarity between a candidate property and the second component of the simile. As before, for properties, context vectors are used. This method is referred to as: $\mathbf{EMB}_1$.

6.3.1.3 Similarity with composite simile vector

For a given event and vehicle, a composite simile vector is created by performing element-wise addition of the vectors for the event and the vehicle, and then cosine similarity with each candidate property is calculated. This method is referred to as: $\mathbf{EMB}_2$.

For example, for <PERSON, talk, robot>, the vectors for "talk" and "robot" are used to create a composite vector, and the similarity between the resulting vector and a candidate property's context vector is used as the ranking criteria. The intuition here is to capture what is common in the context distribution (Mikolov et al., 2013) of "robot" and "talk", and the context vector of a suitable property should have strong similarity with the resulting vector.

Mitchell and Lapata (2008) discussed that a composite vector can be created using element-wise addition or multiplication. For this work, element-wise addition was chosen over multiplication because in this problem setting, addition worked better on the development data.

6.3.2 Results for Candidate Reranking

Table 6.3 presents MRR results after the initially generated candidates are reranked using the influence of the second simile component. For comparison, the MRR results from Table 6.2 are also presented in the first column (**Orig**).

Influence from the second simile component assessed with PMI and EMB_1 improved the MRR scores for some candidate generation methods (e.g., Predicate ADJ and Modifying

Table 6.3: MRR scores for candidate reranking methods using second simile component.

Ranking Method	**Orig**	**PMI**	**EMB$_1$**	**EMB$_2$**
Candidates Generated from Vehicle				
Modifying ADJ	.16	.22	.19	**.24**
Predicate ADJ	.11	.16	.14	**.22**
Explicit Property	.23	.25	.23	**.28**
Dictionary Def.	.22	.21	.20	**.25**
Word Embedding	.15	.19	.20	**.21**
Candidates Generated from Event				
Modifying ADV	.13	.10	.13	**.19**
Explicit Property	.16	**.18**	**.18**	**.18**
Word Embedding	.16	.11	.14	**.18**

ADJ with vehicle candidate source), but did not for others (e.g., Modifying ADV). However, using the composite word embedding vector (EMB$_2$) to capture the common aspects in the context distributions of the event and vehicle consistently improved MRR for all candidate generation methods. Consequently, the composite word embedding vector-based method is used as the final ranking method for each set of candidate properties.

6.4 Aggregated Ranking and Results

Finally, all of the properties produced by the various candidate generation methods are considered in an aggregated ranking. There are many different ways of aggregating individual ranked lists of items. These methods can be broadly categorized into 1) score-based methods and 2) order-based (also known as positional or rank-based) methods (Akritidis et al., 2011). The score-based methods (e.g., Vogt and Cottrell, 1999) utilize scores associated with each ranked item in individual ranked lists. One of the shortcomings of these methods is that the rank scores in the individual lists may not always be directly comparable with each other, as is the case in this work. On the other hand, the positional methods (e.g., Dwork et al., 2001; Klementiev et al., 2008; Lebanon and Lafferty, 2002; Liu et al., 2007) use relative order or position of the items in each ranked list. Akritidis et al. (2011) noted that performance of these two families of rank aggregation techniques are generally comparable, while the latter tends to be more computationally efficient and most techniques can be implemented in linear time with respect to the number and size of the ranked lists.

In this research, to produce an aggregated ranking of all candidate properties, I use a positional method that calculates the harmonic mean of the ranks of a property from the individual candidate generation methods. This approach rewards properties that have a

consistently high ranking across different methods. Table 6.2 showed that the candidate generation methods produce complementary sets of properties and coverage is highest when all of them are used together.

For comparison, results are also shown for 1) the best individual candidate generation method, which uses explicit property extraction patterns to generate candidate properties using the vehicle component of a simile and ranks the properties by conditional probability, and 2) a voting method where a candidate property is ranked based on how many different methods generated it. Ties are broken by frequency of a candidate in the Twitter corpus.

The final results use two gold standard property sets: (1) *Gd* (Gold): uses the set of properties from the human annotators, and (2) *Gd+WN* expands Gold with WordNet synsets (words in the same synset of a gold property are added) and WordNet's "similar to" relation (words that are connected to a gold property by this relation are added). The reason for using *Gd+WN* is to include synonyms of a gold property that would otherwise be considered wrong (e.g., if a human annotator said "beautiful" and the system said "pretty").

The MRR columns in Table 6.4 present MRR results for the final ranking. The results show that with both *Gd* and *Gd+WN*, the aggregated ranking using harmonic mean yields much better MRR results than the best individual method and also better than the Voted method, yielding the highest MRR scores: .33 with *Gd* and .41 with *Gd+WN*.

The MRR metric rewards the rank of the first acceptable property, with diminishing returns for having a low rank. But it does not directly tell us how high the correct property is ranked among all candidate properties. To shed light on this, the *Top 1* and *Top 5* columns of Table 6.4 present the percentage of similes for which an acceptable property was ranked #1 (Top 1) or within the Top 5. The aggregate ranking scheme ranks an acceptable property in the Top 1 position for 27% of the similes based on *Gd+WN*, and identifies an acceptable property within the Top 5 positions for 58% of all similes. This shows that although the property inference framework is able to rank an acceptable property reasonably high for a good percentage of similes, for only about one-fourth of the similes in the data set a property is ranked at the top position, leaving much room for improvement.

Table 6.4: Aggregated ranking results.

	MRR		**Top 1**		**Top 5**	
	Gd	Gd + WN	Gd	Gd + WN	Gd	Gd + WN
Explicit Property (Source: Vehicle)	.23	.30	16%	22%	32%	41%
Voted	.25	.35	14%	21%	36%	52%
Mean	**.33**	**.41**	**21%**	**27%**	**46%**	**58%**

For the previous evaluations, any property given by the annotators is deemed correct, and any consensus that the annotators may have had is not accounted for. To address this, I also conducted an evaluation using only properties agreed upon by multiple human annotators. WordNet synsets and its "similar to" relation are also used in determining consensus. For example, if for the simile <PERSON, look, princess>, one of the annotators provided *beautiful* as the gold property, and a second annotator provided *pretty*, since these two properties are linked to each other by the "similar to" relation in WordNet, they are considered as a single property agreed upon by multiple annotators.

Table 6.5 shows statistics about the properties with respect to their annotation consensus. Row (a) presents the total number of properties agreed upon by at least K annotators across all similes in the data set. The total number of properties reduces from 2,447 (all properties) to 54 (properties that all seven annotators generated). Row (b) shows the number of properties in *Gd+WN*. When synonyms for the properties are considered, the number of properties increases nearly three times with $K = 1$ (from 2,447 to 7,241) and up to nearly 10 times with $K = 7$ (from 54 to 549). In rows (a) and (b), the sudden decrease in number of properties from $K = 1$ to $K = 2$ is likely because when no consensus is required, this results in a diverse set of properties, some of which may not be common or popular interpretations of the respective similes. With consensus, the property sets become more restricted and only contain properties that are the most common interpretations of these similes. As the number of required annotators increases, some similes no longer have any properties that meet the consensus criteria, and row (c) shows the number of remaining similes that have properties meeting the consensus criteria. Row (d) shows the average number of properties per remaining simile with respect to annotator consensus. From $K = 1$ to $K = 2$, the average number of properties (*Gd*) per remaining simile drops drastically from 9.84 to 3.30, but remains within the range of 2.50 to 2.74 properties from $K = 2$ to $K = 7$.

To further explore the issue, I created different "gold standard" data sets where each data set contains only the similes having properties agreed upon by K annotators (ranging

Table 6.5: Statistics for gold standard properties having annotation consensus.

	Annotator Consensus						
	≥1	**≥2**	**≥3**	**≥4**	**≥5**	**≥6**	**≥7**
(a) # of Properties in *Gd*	2,447	821	516	323	211	115	54
(b) # of Properties in *Gd+WN*	7,241	4,025	3,045	2,152	1,670	1,080	549
(c) # of Remaining Similes	641	588	418	252	136	67	27
(d) # of Properties (*Gd*) per Remaining Simile	9.84	3.30	2.51	2.42	2.54	2.70	2.74

from one to seven). So, the similes in these data sets are associated with only the properties that had annotation consensus. Figure 6.3 tracks the property ranking results for these data sets and compares the results with the best individual candidate generation method.

Figure 6.3 shows that for all degrees of consensus, the aggregated ranking is consistently better than the method that uses the explicit property extraction patterns, which was the best individual candidate generation method. The MRR score initially decreases as the number of required annotators increases from one to two, which is probably because two-thirds of the properties in *Gd* are discarded so the average number of properties per simile drops to one-third of the original, which resulted in about 44% reduction of the number of properties in *Gd+WN*. With higher consensus, MRR gradually increases. When more people agree upon a property, the property is likely to have strong salience with respect to the simile components, reflected in stronger statistical association with the component words in text, so comparatively easier to identify by the automated methods.

6.5 Analysis and Discussion

The gold standard property collection confirmed that some similes have many plausible interpretations while others do not. This can potentially contribute to the difficulty of implicit property inference. Utsumi and Kuwabara (2005) introduced the notion of "interpretive diversity" which they referred to as the richness of the figurative meaning of a comparison. They claimed that "interpretive diversity" depends on two factors: 1) the number of features (or predicates) involved in the meaning and 2) the salience distribution of

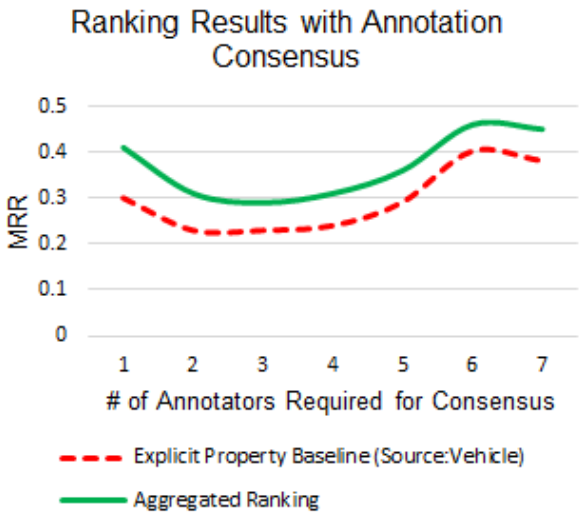

Figure 6.3: Ranking results tracked by annotation consensus with *Gd+WN* gold standard.

those features. They hypothesized that similes with more diversity in the inferred property tend to be more metaphorical, and in these cases the salience values of the properties are more uniform.

They illustrated the case of interpretive diversity with the similes *"Deserts are like ovens"* and *"History is like footprints"*. *"Deserts are like ovens"* conveys one highly salient meaning, which is, "They are burning hot", and relatively less salient meanings such as "They are dry" or "Their temperature is greatly changed". They claimed that these interpretations can be seen as less rich or less diverse. This is because one of the interpretations is likely to be a lot more preferred than the others. On the other hand, *"History is like footprints"* can be considered as having a highly rich, diverse interpretation because many equally salient meanings such as "It remains behind", "It is a thing of the past", and "It is a living proof" are contained in the figurative interpretation. To assess interpretive diversity of a simile, they used Shannon's entropy.

To explore my hypothesis regarding difficulties associated with property inference, I adopted their definition of interpretive diversity as the diversity of implicit properties in a simile (comparable to features/predicates/meanings in the original definition), and measured interpretive diversity using Shannon's entropy. But because the gold standard implicit properties in my data sets are often synonyms of each other, properties that are synonyms first need to be clustered or grouped together. For any simile in the data set, when a property appears in the WordNet synset of another property, or if two properties are connected by the WordNet "similar to" relation, the properties are grouped to form property clusters. So each property cluster represents a set of words that are synonyms of each other. The frequency statistics of individual words in a cluster are then aggregated.

To illustrate this with an example, let's consider that for the simile <room, feel, Antarctica>, the annotated properties are *cold* and *frigid* by annotator#1, *cold* and *big* by annotator#2, and *cold* and *white* by annotator#3. So, the frequency (within parentheses) of each individual property is: *cold* (3), *frigid* (1), *big* (1), and *white* (1). Now, from the WordNet's synsets and "similar to" relation, *cold* and *frigid* are synonyms. Consequently, the following property clusters are formed: {cold, frigid}, {big}, and {white}. The aggregated frequency for the clusters are {cold, frigid} (3 + 1 = 4), {big} (1), and {white} (1). With the aggregated frequencies of the property clusters, interpretive diversity of each simile is finally measured using Shannon's entropy using the following:

$$H(X) = -\sum_{x \in X} P(x) \log_2 P(x)$$

Here, X is the random variable representing property clusters of a simile, and $P(x)$ is the probability of a cluster $x \in X$, calculated by,

$$P(x) = \frac{aggregated\_frequency(x)}{\sum_{x \in X} aggregated\_frequency(x)}$$

Figure 6.4 shows the entropy curve after the 641 similes of the data set are sorted by the entropy values of their property clusters. Based on changes in the slope of the curve, the similes can be divided into three classes for the sake of analysis. The first class contains 100 similes with entropy values $\geq$ 3.33 and can be considered as similes with high interpretive diversity. The next 400 Similes with entropy values $\geq$ 2.42 and <3.33 are taken as similes with medium interpretive diversity. And finally, the remaining 141 similes are similes with low interpretive diversity with entropy values <2.42.

Table 6.6 presents examples of similes in each category. Each property cluster having more than one property in the cluster is enclosed within curly braces, and the aggregated frequency of a cluster or frequency of a property appearing more than once are presented within parentheses.

High interpretive diversity is clearly demonstrated by <PERSON, act, mom>, showing properties with many different characteristics attributed to mom.[7] Note that the properties contain both positive (e.g., friendly, loving) and negative (scolding, annoying)

[7]The spelling mistake in "nuturing" naturally occurred during the annotation.

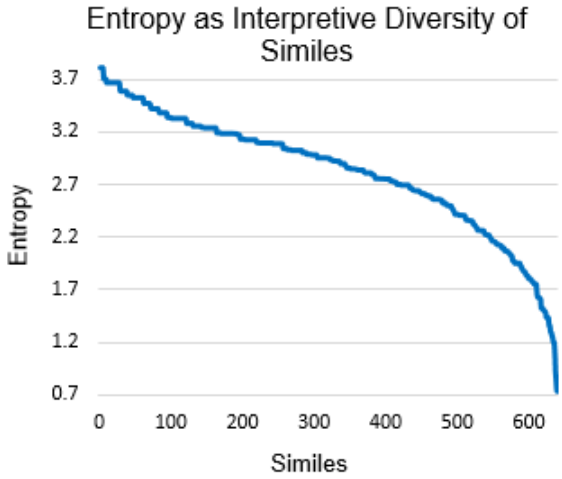

Figure 6.4: Entropy as interpretive diversity of similes.

Table 6.6: Similes with different levels of interpretive diversity. Property clusters are enclosed with curly braces. Aggregated frequencies are presented within parentheses. The properties are from the gold standard.

Simile	Gold Properties
High Interpretive Diversity	
<PERSON, act, mom>	bossy (2), friendly, nuturing, overbearing, loving, scolding, caring, hovers, strict, protective, cleans, nurturing, annoying
<PERSON, act, baby>	{childish,immature,young} (4), crying (2), whine, silly, cry, dependent, needy, pouting, whiny, weak
Medium Interpretive Diversity	
<PERSON, look, robot>	stiff (5), jointed, stoic, blank, expressionless, mechanical, inhuman, dull, uneasy
<girl, be, butterfly>	{beautiful,pretty} (4), free (2), delicate (2), graceful (2), fluttering, floating, happy, flowy
Low Interpretive Diversity	
<PERSON, act, clown>	{goofy,ridiculous,silly} (5), {amusing,comical,funny} (5), stupid, degrading, disruptive, childish
<throat, feel, sandpaper>	{rough,scratchy} (9), coarse (2), raspy, sore, dry

attributes. Both similes with high interpretive diversity show that their property frequencies are relatively uniformly distributed. On the other side of the spectrum are similes with low interpretive diversity, as exemplified by <PERSON, act, clown> and <throat, feel, sandpaper>. The simile <PERSON, act, clown> has two clusters of properties with a high aggregated frequency of five by each, and <throat, feel, sandpaper> has one cluster of properties with an even higher aggregated frequency of nine.

Table 6.7 presents property ranking results for these three different classes of similes using *Gd+WN*. The results show that it is much harder to infer the implicit property in similes with high interpretive diversity, demonstrated by a .19 difference in MRR score from high to low. This trend is also consistent in the percentage of similes for which the system ranks a plausible property at the topmost position (Top 1) or within the Top 5. It is possible that with low interpretive diversity, statistical associations between a property and simile components are stronger, and so more easily discovered by the candidate generation and

Table 6.7: Results for different subsets of similes divided by interpretive diversity, using *Gd+WN* properties.

Diversity	**High**	**Medium**	**Low**
MRR	.31	.40	.50
Top 1	15%	26%	37%
Top 5	47%	57%	66%

ranking methods.

Table 6.8 presents the same examples from Table 6.6, but with properties that are inferred by the system. For the two similes with high interpretive diversity, the system was unable to infer a gold property within the top 5 for <PERSON, act, mom>, and the gold properties identified for <PERSON, act, baby> were ranked 4th and 5th. For this second example, properties such as "cute" and "adorable" are highly salient properties for baby, but they were not generated as properties of this simile by the human annotators.

For the two similes with medium interpretive diversity, the system was able to infer a gold property for both cases, but for <PERSON, look, robot>, the gold property is ranked lower, and for <girl, be, butterfly>, the gold properties are ranked high. "Emotionless" might be considered as an acceptable property, but it was not present in the gold standard. For the two similes with low interpretive diversity, the system was able to infer multiple gold properties and was also able to rank them in the top positions.

6.6 Improving Affective Polarity Recognition Using Inferred Properties

Being able to infer implicit properties in similes provides additional information that was not available before for affective polarity classification. For example, the supervised classification method presented in Chapter 5 derived lexical, semantic, and sentiment features from all of the observable components of a simile, i.e., the tenor, event, vehicle, and the explicit properties that have been previously associated with the vehicle in comparisons. The inferred implicit properties, on the other hand, allow deeper understanding of the basis of the comparison and meaning of the similes when there is no property explicitly mentioned.

In this section, I create additional features from the inferred implicit properties and show

Table 6.8: Example output of the inferred properties (in ranked order from left to right). Properties from *Gd+WN* are in boldface.

Simile	**Inferred Properties**
High Interpretive Diversity	
<PERSON, act, mom>	funny, cooking, cool, dramatic, hot
<PERSON, act, baby>	cute, adorable, unborn, **immature**, **newborn**
Medium Interpretive Diversity	
<PERSON, look, robot>	human, emotionless, giant, cool, **mechanical**
<girl, be, butterfly>	**beautiful**, **free**, cute, weird, crazy
Low Interpretive Diversity	
<PERSON, act, clown>	**funny**, **stupid**, **silly**, real, **goofy**
<throat, feel, sandpaper>	**dry**, **rough**, stiff, heavy, smooth

that they can be used to improve affective polarity recognition. The inferred properties of a simile themselves are new lexical information about the similes, so they can be added to the supervised classifiers as additional lexical features. Sometimes the inferred properties are positive/negative words that can be identified using existing sentiment lexicons. The polarity information about the properties can act as priors, and can be added to the supervised classifiers as additional sentiment features for affective polarity classification.

6.6.1 Using Implicit Properties as Additional Features

To use the inferred properties as additional features for the supervised affective polarity classifier, the following features are used:

6.6.1.1 Inferred Properties as Lexical Features

Using the property inference framework, the top five properties from the final aggregate ranking are extracted. For each simile instance, the extracted top properties are supplied to the supervised classifier as five additional lexical features. The features are binary indicating the presence or absence of an inferred property for the simile.

6.6.1.2 Property Sentiment Features

Two binary features (one for positive sentiment and one for negative sentiment) represent whether there is a positive sentiment word among the top 5 inferred properties and whether there is a negative sentiment word among the top five inferred properties. To determine the sentiment, the combined AFINN sentiment and MPQA subjectivity lexicons are used.

6.6.2 Affective Polarity Classification Results for Manually Annotated Training Data

Table 6.9 presents the results for training with manually annotated data when the additional features derived from inferred properties are added to the original feature set. For comparison, the best results obtained using the original feature set are also presented at the top row of the table.

The lower section of Table 6.9 shows the results after adding the implicit property features. When the inferred properties are added as lexical features, classification results for the positive polarity class improves with 3% additional precision, and 5% additional recall, resulting in a 3% F1-score improvement over the previous best results. On the other hand, for the negative polarity class, a precision-recall trade-off (1% precision gain but 2% recall drop) resulted in a similar F1-score as before. Results for the neutral category improves by an additional 4% F1-score. This could be because for similes with neutral

Table 6.9: Results with implicit property features for manually annotated training data (P = Precision, R = Recall, F = F1-score).

	Positive			**Negative**			**Neutral**		
	P	R	F	P	R	F	P	R	F
Previous Results									
(a) Unigrams + Other Lexical + Semantic + Sentiment (excluding inferred property features)	75	60	67	77	**79**	78	25	40	31
With Features Derived from Implicit Properties									
(a) + (b) Property Lexical Features	**78**	**65**	**70**	78	77	78	**28**	**46**	**35**
(a) + (c) Property Sentiment Features	77	61	68	**79**	**79**	**79**	25	45	32
(a) + (b) + (c)	77	**65**	**70**	**79**	**79**	**79**	**28**	45	34

polarity, the inferred properties are less likely to have polarity themselves. If similar set of properties are present for the neutral similes in the training data, with this new information, the classifiers no longer labeled some of the neutral similes as positive or negative, which resulted in more correctly classified neutral similes than before.

When the property sentiment features are added to the original feature set, they show improvements over all three categories. With these features, 2% precision and 1% recall gain resulted in a 1% F1-score improvement for the positive class, and a 2% gain in precision with no recall drop resulted in a 1% F1-score improvement for the negative class.

When both types of features are added together to the original feature set, the final row in Table 6.9 shows that the best performance can be achieved for both the positive and negative classes. The combined feature set yields the highest 70% F1-score and 79% F1-score for the positive and negative classes, respectively, suggesting that the property information is valuable for recognizing the affective polarity in similes.

I have also conducted statistical significance test using paired bootstrap (Efron and Tibshirani, 1994; Berg-Kirkpatrick et al., 2012) to determine if the improvements are statistically significant. For the positive polarity class, the recall and F1-score improvements (using both lexical and sentiment property features) over the previous best results (without the property features) have been found to be statistically significant at 95% confidence level (p =0.03 for both). For the negative polarity class, the precision improvement (using both lexical and sentiment property features) over the previous best results (without the property features) have been found to be statistically significant at 90% confidence level (p =0.06).

6.6.3 Affective Polarity Classification Results for Automatically Labeled Training Data

Table 6.10 presents results with the added features when using the automatically labeled training instances. With both property lexical features and property sentiment features,

Table 6.10: Results with implicit property features for automatically labeled training data (P = Precision, R = Recall, F = F1-score).

	Positive			**Negative**			**Neutral**		
	P	R	F	P	R	F	P	R	F
Previous Results									
(a) SVM with labeled data from all sources (w/o features from inferred properties)	**61**	61	61	70	52	60	12	**24**	16
Features Derived from Implicit Properties									
(a) + (b) Property Lexical Features	60	**64**	**62**	**72**	**55**	**62**	**13**	23	**17**
(a) + (c) Property Sentiment Features	**61**	62	**62**	71	52	60	12	**24**	16
(a) + (b) + (c)	58	**64**	61	**72**	54	**62**	**13**	22	16

recall improves for the positive class (statistically significant at 90% confidence level, $p = 0.06$, using the paired bootstrap method). However, there is a precision-recall trade-off and the final F1-score remains similar as before.

On the other hand, for the negative class, the new lexical property features improve both precision and recall, resulting in an additional 2% F1-score improvement. With the property sentiment features, precision improves slightly by 1%, but the recall and F1-score remain similar as before. When both types of features are added together, this results in a consistent 2% precision, recall, and F1-score improvement (statistically significant at 90% confidence level, p =0.098, using the paired bootstrap method) over the previous best results.

The results for the neutral categories remained same as before, probably due to the limitations of the methods for being able to automatically identify neutral instances to provide good-quality neutral training instances for the classifiers. So, the additional information about the inferred properties did not improve classification results for the neutral class when automatically labeled training data are used.

6.7 Chapter Summary

In this chapter, I have presented a framework for inferring implicit properties in similes that breaks down the problem into three sub-tasks. First, candidate properties are generated from the event and vehicle components individually and the properties are ranked using some initial ranking criteria. The generated candidates are then reranked using the influence of a complementary component which was not the original source, so that compatibility with the complementary component can help to rank a plausible property higher. Finally, individual ranks from multiple methods are aggregated using the harmonic mean.

The candidate generation methods use explicit property extraction patterns, measures

of statistical associations, dictionary definitions, and word embedding vector similarity to generate candidate properties using the vehicle and event components of a simile. The candidate generation methods are able to generate at least 10-20 candidate properties for most similes in the data set, and together are able to generate at least one acceptable property for 86% of the similes in the data set.

The best method reranks the generated candidates using word embedding vector similarity with a composite simile vector, and aggregates the ranks of individual methods using the harmonic mean. The final results with aggregate ranking show substantial improvements over the best individual candidate generation method as well as a voted method.

Performance of implicit property inference is evaluated with respect to interpretive diversity of similes. The presented analysis suggests that similes with high interpretive diversity, i.e., similes that have many different implicit properties are harder to infer automatically than similes with low interpretive diversity.

The inferred properties are used to improve affective polarity classification. The inferred properties can be easily incorporated into the supervised classification framework as additional lexical features, and additional sentiment features can be derived from them. Experiments show that the new features consistently improve affective polarity classification performance for both manually annotated data and automatically labeled training data.

The presented framework is able rank a gold property within the top give ranked properties for many similes, but not always at the top position, leaving room for improvement. The presented methods also did not use the tenor component of a simile, which for some similes can be a valuable source of information. Being able to identify the similes where the tenor will be informative for property inference can be a potential improvement for future work.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

In this dissertation, I have presented research on acquiring knowledge for affective state recognition in social media. As the affective knowledge, the research addresses two types: 1) hashtag indicators of emotions consisting of emotion hashtags, emotion hashtag patterns, and emotion phrases; and 2) affective understanding of similes. Both hashtags and similes are common in Twitter, a popular microblogging social media platform. People frequently express their feelings in tweets, making Twitter an ideal source of data for this research. In this chapter, I summarize the research contributions and findings, and also discuss avenues for future work that can stem from this research.

7.1 Claims and Research Contributions Revisited

In this section, I will discuss to what extent the conducted research supports the claims I made at the beginning of this dissertation and the underlying research contributions of the claims.

> *Claim#1: Hashtag indicators of emotions can be automatically learned from tweets using a bootstrapped learning framework.*

In Chapter 3, I presented a bootstrapped learning framework that can successfully learn emotion hashtags with respect to five emotion categories: AFFECTION, ANGER/RAGE, FEAR/ANXIETY, JOY, and SADNESS/DISAPPOINTMENT. Starting with only five emotion hashtags per category, the collection of seed emotion hashtags is grown into a much bigger collection containing hashtags ranging from 260 (for AFFECTION) to 940 (for ANGER/RAGE) hashtags. The creation of the initial seed hashtags requires minimal human effort, and the rest of the bootstrapped learning algorithm runs automatically. The optimum number of hashtags for each emotion category is decided on at the end of the bootstrapped learning by evaluating how well they can classify emotion in a development data set.

The quality of the learned hashtags is reflected in the evaluation task of tweet emotion classification (Table 3.6, Section 3.5.2) where the hashtags are used to predict the emotion of a tweet. Precision ranges from 60%-82%, demonstrating that the learned hashtags are good

quality indicators of the emotion categories. Recall ranges from 13%-37%, indicating that the learned emotion hashtags often appear in emotion tweets, but they are not sufficient to recognize all such tweets.

The bootstrapped learning framework is then extended to also learn emotion hashtag patterns. The learned hashtag patterns range from 270 (for JOY) to 970 (for FEAR/ANXIETY). They improve emotion classification recall (Table 3.6 in Section 3.5.2) by an additional +14% for AFFECTION and +5% for FEAR/ANXIETY over the use of specific emotion hashtags, demonstrating that the patterns can generalize beyond specific hashtags. Recall did not improve for JOY and SADNESS/DISAPPOINTMENT, and had marginal improvement for ANGER/RAGE (+1%). For these emotion categories, the optimum number of emotion hashtag patterns was fewer than the optimum number of emotion hashtags, based on development data. Finding better ways to learn when to stop generating hashtags can be an important avenue for future work.

At the end of the bootstrapped learning, emotion phrases are harvested by expanding the learned emotion hashtags and patterns. The emotion phrases are then used to train additional emotion classifiers that use the surrounding context words of the emotion phrases as features. The learned emotion phrases (Table 3.7 in Section 3.5.3) do not always work well when they are directly used for emotion classification. Precision scores range from 19% to 53% for the emotion phrases learned from hashtags, and from 33% to 62% for the emotion phrases learned from hashtag patterns. But training context-based classifiers improves precision. With the flexible context model, precision scores for the classifiers range from 48% to 69% across the emotion categories, demonstrating that the context-based emotion classifiers trained using the emotion phrases have prediction value.

The learned hashtag indicators of emotions can be best used in a hybrid approach that labels a tweet with an emotion when EITHER the supervised classifier (with unigrams and prediction probability features from the phrase context-based classifiers) predicts the emotion for the tweet, OR a learned emotion hashtag or hashtag pattern is present in the tweet. The hybrid approach improves macro-average F1-score by +13% over the baseline supervised classifiers that use only unigram features. This demonstrates that the bootstrapping framework can successfully learn good-quality hashtag indicators of emotions.

> *Claim#2: Affective interpretation of similes can be automatically achieved by affective polarity classification of similes and by inferring the implicit properties of open similes.*

In Chapter 5, I presented supervised classifiers that extract features reflecting lexical, semantic, and sentiment properties of simile components. The classifiers assign a positive,

negative, or neutral label to a simile indicating its affective polarity. Using training instances that are manually labeled, the classifiers improve affective polarity classification performance by +20% F1-score for positive polarity and by +29% F1-score for negative polarity, over the best results achieved using existing sentiment resources (Table 5.5 in Section 5.5).

Using training instances that are automatically labeled, the supervised classifiers with the full feature set achieve up to 61% and 60% F1-score for the positive and negative polarity classes, respectively (Table 5.5 in Section 5.5). The results with the automatically labeled training instances are not as high as the results with the manually labeled data, but they are still substantially better than the results from the baseline methods that use existing sentiment resources to recognize the affective polarity of a simile.

In Chapter 5, I have also presented the design for a framework to automatically infer implicit properties in similes. The problem is decomposed into 3 subtasks. First, a variety of candidate generation methods are used to generate candidate properties individually from the *vehicle* and *event* components of a simile. The candidate generation methods use syntactic extraction patterns, statistical association, dictionary definitions, and word embedding vector similarity to generate and rank initial lists of candidate properties. Top candidate properties are then reranked by the influence of a complementary component to improve the ranking. The ranks for each property from individual methods are then aggregated to generate a final ranked list of candidate properties.

The reranking results demonstrate that considering influence from multiple simile components makes a positive impact on improving candidate property ranking. The final results show that the final aggregate ranking improves property inference by 0.10 MRR over the best candidate generation baseline method that use explicit property extraction patterns, and by 0.08 MRR over a voted method by the individual candidate generation methods (Table 6.4 in Section 6.4). When the gold implicit properties are expanded with WordNet, the ranking performance improves up to 0.41 MRR. Additional analysis shows that it is harder to infer a property for similes that have high interpretive diversity, but for similes with low interpretive diversity, the presented method performs even better, achieving up to 0.50 MRR.

When the inferred implicit properties are added as additional lexical features to the supervised classifiers, affective polarity recognition in similes is further improved. The additional features improve precision by 3% and recall by 5%, resulting in a 3% F1-score improvement for positive polarity, when manually annotated training instances are used.

Classification results for the negative polarity class showed a precision-recall trade-off which yielded similar F1-score results as before.

When automatically labeled training instances are used, recall improves by 3% for the positive polarity class, resulting in an additional 1% F1-score improvement. For the negative polarity class, both precision and recall improve by 3%, resulting in an additional 2% F1-score improvement, demonstrating that the knowledge of the inferred implicit properties contribute to the affective interpretation of similes.

7.2 Future Work Directions

The research presented in this dissertation opens up a number of avenues that can be pursued in future work. They can be divided into two broad categories: 1) improvement scopes for this work, and 2) novel application areas.

7.2.1 Improvement Scopes

Some aspects of this research can be explored more in future work to address their current limitations. Some of these improvement scopes are discussed below.

7.2.1.1 Stopping Condition during Iterative Learning

The bootstrapping framework for learning hashtag indicators of emotions did not have a stopping criteria; rather the learning algorithm was run for a fixed number of iterations. At the end, optimum sizes of the learned lists of hashtag indicators were determined by evaluating their performance on a tuning set using the lexicon look-up method with incremental list sizes of the hashtag indicators.

Because the optimum sizes are determined post hoc, noise can get introduced during learning without being detected, which can potentially steer the learning in the wrong direction. Noisy hashtags will result in both noisy positive training instances (for the emotion categories they are learned in) and negative training instances (for the rest of the emotion categories) for the emotion classifiers.

One way to address this can be by adopting methods similar to the one introduced by McIntosh and Curran (2009). McIntosh and Curran (2009) introduced the notion of semantic drift detection during bootstrapped learning of semantic lexicons. Using distributional similarity, they determined when learned words at later stages semantically differed from the seed terms. In the case of hashtag indicators of emotions, by determining when the contexts where newly learned hashtag indicators start to substantially differ from the contexts where the seed hashtags appear, suitable stopping conditions may be developed.

This can potentially improve the quality of the hashtag indicators that are learned during later iterations of the bootstrapping algorithm.

7.2.1.2 Phrase Context Modeling

The prediction probability features from the phrase context-based classifiers described in this research gain marginal improvements when added to the supervised classifiers that use unigram features. The phrase context-based classifiers themselves are not very strong as precision ranges from 48% to 69% with low recall, under the flexible context model. This leaves ample room for improvement in future work.

One potential step forward can be to model the context of an emotion phrase targeting relevant aspects of the problem. For example, some of the problematic cases are related to hypothetical statements or future tense cases. In hypothetical statements, the writer puts forth a proposition which is often conditioned upon other circumstances. For example, the statement *"I would have been really angry at mom if she did not let me play"* does not indicate that the writer is actually feeling the emotion ANGER, despite the presence of the phrase "angry at". Although the current model would consider the context words "would", "have", "felt", etc., the information regarding if the context is hypothetical can be modeled more explicitly.

Similarly, statements with future tense do not normally represent the current emotional state of the writer. For example, *"I will be happy tomorrow when the exam is finally over"* does not indicate that the writer feels the emotion JOY despite the presence of the word "happy". Rather, the writer is stating a future circumstance that has not happened yet. Information such as this can also be explicitly modeled.

It is also important to make sense of who feels an emotion. In the case of hashtags, the emotion hashtags most commonly represent the writer's emotional state, whereas the emotion phrases derived from the same hashtags may have a different subject in the body of the tweet. For example, *"he is super excited"* does not refer to the writer's excitement, and needs to be modeled accordingly.

The topics in the context can also be explored. For example, the phrase *"you rock"* is more strongly associated with the emotion AFFECTION when there is a mention of a family member in the context. Similarly, mention of an insect can have a strong association with FEAR/ANXIETY or a mention of a festival or holiday with JOY, etc. These potentially relevant aspects of the contexts can be specifically targeted, and modeling the context of an emotion phrase accordingly can be a promising direction for future work.

7.2.1.3 Hashtag Emotion Indicators for Other Languages

The research presented in this dissertation on learning hashtag indicators of emotions is performed for the English language, but the methods and algorithms introduced are not language-specific. So, they can potentially be adopted for learning hashtag indicators of emotions for other natural languages too (e.g., Spanish, Portuguese, etc.), thus improving coverage of the learned affective knowledge across multiple languages. The bootstrapping algorithm requires a small number of seed emotion hashtags, which would take little manual effort to acquire for the target natural language. During learning, the classifiers use unigrams and bigrams in tweets as features for training emotion classifiers. With sufficient collection of seed hashtag labeled tweets and unlabeled tweets written in the target language, this can also be achieved.

One potential challenge would be to determine if emotion hashtags are also frequently used in the target language that is not English, as usage practice may vary from language to language. Different techniques may also be required to acquire N-grams in some target languages; for example, Chinese sentences are written as continuous character strings (Nie et al., 2000). It is also important to supply the emotion classifiers with sufficient seed hashtag labeled tweets, which may be challenging to acquire for some languages if Twitter or social media in general is not as popular for people who are native speakers of these languages.

7.2.1.4 Exploring Influence of Tenors

For inferring implicit properties in similes in this research, the influence of the *event* and *vehicle* has been explored, which leaves the experimental investigation of the *tenor*'s influence for future work. In the data set, nearly 60% of the similes have pronouns as tenors, which are semantically weak, so do not provide useful information without performing coreference resolution. But the other 40% of similes do not have pronominal tenors, so they can be potentially used for implicit property inference.

In Section 4.2.4, the example *"my room feels like Antarctica"* has been discussed to demonstrate that many properties generated from "Antarctica" can also be compatible with "room", in which case the tenor is not able to limit the inference space. But there are also other examples, such as *"time be like river"*, where the tenor can play an important role in eliminating some of the properties that come from river but do not fit in the context of the simile. Identifying these cases where the tenors can contribute to implicit property inference should be explored in future work.

7.2.1.5 Exploring Rank Aggregation Techniques

During ranking of candidate implicit properties for a simile, this research aggregates ranks from different methods using harmonic mean. This is a positional method as the ranking criteria is a reflection of different positions of a property in individual ranked lists. While computationally efficient, positional methods can further benefit from addressing certain limitations. For example, they do not account for individual rank scores, the sizes of the ranked lists may vary, relative differences of the positions in a ranked list can be factored in, etc.

There are dedicated research works that directly address the rank aggregation problem. For example, rank aggregation by creating initial ranked lists of items using Markov Chain optimized with Local Kemenization (Dwork et al., 2001), factoring in distance between ranks (Klementiev et al., 2008), using conditional probability models on permutations (Lebanon and Lafferty, 2002), using supervised rank aggregation (Liu et al., 2007), etc. These methods can be actively explored in future work to investigate if they can produce improved rankings over the aggregate ranking method used in this research.

7.2.2 Novel Application Areas

My research also opens up a number of novel application areas that can be pursued in future work. These areas may correspond to further applications of the methods used in this research or the applications of the types of affective knowledge acquired in this research. Some of these application areas are discussed in the following sections.

7.2.2.1 Discovering Hashtags Associated with Events

Expressing an emotion is a popular use of hashtags in social media. But there are also other uses of hashtags. It is a common practice among Twitter users to create hashtags during ongoing events (also during social or political movements), so that anyone can search for these hashtags to know about other people's thoughts/opinions as well as any update involving these situations. For example, #CoatHangerRebellion was created to represent a protest against an anti-abortion law in Poland, #BlackLivesMatter is associated with an activist movement campaigning against violence toward people with African-American racial origin, etc. A promising future work direction could be to use a similar bootstrapped learning framework as used in this research to automatically learn hashtags that are associated with similar types of events and movements.

For the examples above and also for other events that are similar, there are some aspects about the events that may be common. For example, the entities involved in a protest

event may include police, journalists, and government, and the props involved may include banners, placards, etc. As a result, when people describe these events in tweets, there may be lexical and semantic overlap among the text contents of the tweets, making it ideal for training classifiers to recognize events of similar types. Using a small number of seed event hashtags of similar types, the bootstrapped learning framework can be used to learn more hashtags that are created for new events of similar types. Different types of events for which hashtags can be learned may include: civil unrest events (e.g., protests, strikes, demonstrations), natural disasters (e.g., earthquakes, tornadoes, hurricanes, wildfire), or political events (e.g., presidential primaries, presidential elections, political campaigns).

7.2.2.2 Learning Tenor States with Affective Polarity

Inferring an implicit property in a simile allows for acquiring state knowledge about the subject of comparison (i.e., the tenor). For example, by inferring "cold" as the implicit property in the simile *"my room feels like Antarctica"*, state knowledge about the room can be acquired as *"my room is cold"*. In many of the cases, the affective polarity of a simile will be the same as the affective polarity of the tenor's state. My research presented methods for associating affective polarity to a simile. From knowing that *"my room feels like Antarctica"* describes a negative state of the room, the same polarity can also be associated with the state: *"my room is cold"*. Thus the research on recognizing affective polarity of similes and inferring their implicit properties can be combined together to acquire knowledge of state descriptions with affective polarity that are not explicitly described in the similes.

A potential challenge is that there can be some cases where the polarity of an acquired state description may differ from the state described in the simile. For example, *"my hair looks like a poodle"* has a negative polarity as a simile. The most likely implicit property for this simile is "curly". But the state description *"my hair is/looks curly"* is a statement simply describing the state of the hair, and does not have a negative polarity.

One way this can be overcome is by analyzing and identifying when sentiment polarity in the surrounding contexts of these two state descriptions is not consistent. If the simile has a negative polarity, some of the surrounding contexts can be expected to have words with negative sentiment. On the other hand, since *"my hair is/looks curly"* is neutral, the surrounding context words will not consistently have negative sentiment words. The learned states with affective polarity can further enrich the acquired affective knowledge presented in this research.

7.2.2.3 Affective Polarity of Metaphors

A natural extension of the research presented in this dissertation is the use of the learned affective polarity of similes in determining affective polarity of metaphors. Section 4.1.2 discussed how metaphors and similes differ from and relate to each other. There are many cases where a metaphor can be treated as an elliptical simile, especially in case of predicate nominals (e.g., *"he is like a lion in battlefield"* vs. *"he is a lion in battlefield"*). In these cases, the affective polarity of the similes can be aligned with the affective polarity of the corresponding metaphors. But a conversion such as this is not always possible, leaving it as a challenge to identify and treat these cases separately.

There are also cases when a conversion may not be necessary; rather identifying the component terms may be sufficient. For example, identifying the word "zombie" as a vehicle in the statement *"my zombie look after a long day of work made everyone worried"* may be sufficient to determine that a zombie look has a negative polarity. In cases such as these, a metaphor does not need to be explicitly transformed into a simile, but the individual components of the metaphorical comparison can still be aligned to the components of a simile.

Metaphors cover a broader range of figurative comparisons, some of which are similar in construction to similes. A potential future work direction stemming from this research is the exploration of to what extent the knowledge of the affective polarity of similes can be adopted to benefit the understanding of affective polarity of metaphors.

7.2.2.4 Recognizing Sarcastic/Ironic Similes

There is an interesting interplay between sarcasm and similes. Riloff et al. (2013) demonstrated that sarcasm is often created by expressing a positive sentiment toward a negative situation. As many similes have positive or negative affective polarity, people often write similes sarcastically. Sometimes a positive sentiment is explicitly expressed in a simile that has negative polarity, and other times a negative situation or concept is compared with a positive concept. There are also cases where the comparison indicates a property which is the opposite of what is the actual case. Figure 7.1 presents some examples of these cases from actual tweets.

In the first two examples in Figure 7.1, the tenors present concepts with negative polarity: "stinging nettles" and "gun shots". These two concepts are compared with "warm hugs" and "music to ears" having positive polarity, making the comparisons sarcastic. The third example demonstrates the case of an opposite property. The concept "stand" is neither positive nor negative, but it refers to a firm stance that should not change position

Figure 7.1: Examples of tweets with similes and sarcasm.

(figuratively or literally). But by comparing it with a pendulum in strong wind, the writer indicates that the "stand" is not firm, making it a sarcastic comparison. In the last example, the negative situation "nosebleeds" is complemented explicitly with a positive sentiment by saying it is the "best part about winter", indicating sarcasm.

Understanding these sarcastic statements is rooted in understanding the affective polarity and the implicit properties in these similes. Veale and Hao (2007) mentioned that 13% of all simile instances and 20% of all simile types in their data set were ironic, which makes sarcasm a frequent phenomenon among similes. Using the research presented in this dissertation for affective polarity recognition and implicit property inference, a potential future work avenue will be to gain a deeper understanding of similes so that sarcastic/ironic similes can be identified automatically.

7.3 Summary

In this dissertation, I have presented research work on acquiring knowledge for affective state recognition in social media. The major contributions of this research are the novel types of affective knowledge addressed, that were not explored by previous research, as well as the methods and frameworks for acquiring them automatically. The experimental results suggest that the addressed affective knowledge can be successfully learned using the proposed automatic methods. The research presented in this dissertation opens up a

number of promising avenues for future work, including expanded scopes of the research presented in this dissertation and novel application areas that can arise from this work.

APPENDIX A

TWEET EMOTION ANNOTATION GUIDELINES

Task Description

The task requires assigning one or more emotions (**maximum two**) to a tweet written in English, based on the emotion(s) the **writer** felt. The task addresses five emotion classes and and a sixth class representing "other".

Emotion Definitions

1. Affection: a feeling of fondness or tenderness for a person or animal. The most common cases of affection involve the emotional state felt towards family members, friends, loved ones, pets etc. A romantic feeling felt towards any person should be considered affection. However, affection should be different from any "general" liking. For example, liking the president of a country is not affection. Affection can not be shown towards an inanimate object. Liking towards an inanimate object should be considered Joy instead. The class covers adoration, affection, love, fondness, caring, tenderness, compassion.

2. Anger/Rage: a feeling of great annoyance or antagonism as the result of a real or perceived grievance; rage; wrath. The class covers anger, rage, fury, wrath, hostility, ferocity, hate, loathe.

3. Fear/Anxiety: a feeling of distress, or alarm caused by impending danger, pain, etc. Also a state of uneasiness or tension caused by apprehension of future uncertainty, misfortune, etc. The class covers fear, fright, horror, terror, panic, scare, hysteria, nervousness, tenseness, uneasiness, anxiety, apprehension, worry, dread.

4. Joy: a feeling of happiness or contentment; pleasure or satisfaction directed at a person or inanimate object. Liking towards an inanimate object should be considered Joy. The class covers amusement, cheerfulness, gaiety, glee, jolliness, joviality, joy,

delight, enjoyment, gladness, happiness, jubilation, elation, satisfaction, euphoria, enthusiasm, zeal, zest, excitement, thrill, exhilaration, eagerness.

5. SADNESS/DISAPPOINTMENT: the quality or state of feeling sorrow, or the emotional state felt by the failure of an expectation. The feeling of unhappiness at a situation or with someone or something should also be considered SADNESS/DISAPPOINTMENT. The class covers depression, despair, hopelessness, gloom, glumness, sadness, grief, woe, misery, melancholy, disappointment, displeasure, dislike, unhappiness.

6. OTHER: if a tweet contains 1) no emotion, or 2) an emotion that does not clearly fall under the five emotions above, the OTHER label should be used.

Important Note: The annotation should be with respect to the emotion the writer feels, and **not** the emotion felt by someone mentioned in the tweet or by the reader of the tweet.

Examples of Emotion Tweets

(a) "I love my best friend" (AFFECTION)

(b) "Shut your mouth and get the hell out of here #idiot" (ANGER/RAGE)

(c) "#glad to be here today" (JOY)

(d) "there's a spider on my ceiling! someone get it out!!!" (FEAR/ANXIETY)

(e) "miss New York so much! #depressed" (SADNESS/DISAPPOINTMENT)

Examples of Multiple Emotions in Tweets

(a) "My mom is the only person who cares about me #depressed. Thanks for always being there for me mom." (AFFECTION, SADNESS/DISAPPOINTMENT)

Examples of "Other"

(a) "You should always try to be happy" (not felt by the writer → OTHER)

(b) "Going to the #game" (no emotion → OTHER)

(c) "there you are! I have been looking for you!" (none of the five emotions → OTHER)

(d) "My boss #fired me today" → OTHER. (Commonly, such events can induce SADNESS/DISAPPOINTMENT or ANGER/RAGE, but it is not clear from the tweet if the tweeter felt one of these emotions)

Note: The content words and the hashtags should both be taken into account. Sometimes the emotion may be clear from just content or just hashtag, or both together.

Examples of Emotion from Content Words and Hashtag

(a) "#class is starting tomorrow" (content/hashtag express no emotion → OTHER)

(b) "class is starting tomorrow #excited" (emotion is clear from hashtag → JOY)

(c) "you seem #depressed" → OTHER (hashtag may suggest SADNESS/DISAPPOINTMENT, but content makes it clear it is not about the writer)

(d) "I am #depressed" → SADNESS/DISAPPOINTMENT (hashtag and content both suggest the writer is sad)

Comparison of Similar Tweets for Affection and Joy

(a) "I love my mom!" (AFFECTION)

(b) "I love my new laptop!" (JOY)

(c) "My new kitten is so sweet!" (AFFECTION)

(d) "My dog just learned a new trick! yay!" (JOY)

For the tweets below, select the most appropriate emotion, or OTHER. If more than one emotion clearly applies, use a maximum of 2 emotions.

APPENDIX B

SIMILE AFFECTIVE POLARITY ANNOTATION GUIDELINES

Task Description and General Instructions

For this task, you should judge whether a positive or negative sentiment is felt toward the subject of a simile. A simile is a figure of speech comparing two essentially unlike things, such as "Jane swims like a dolphin". Each simile should be assigned one of 4 labels: POSITIVE, NEGATIVE, NEITHER, or INVALID. You should assign the label that typically describes the sentiment toward the subject of the simile (the leftmost word, "Jane" above). For example, dolphins are excellent swimmers, so "Jane swims like a dolphin" is a compliment toward Jane and should be labeled as having a POSITIVE sentiment.

Usually, you should be able to label a simile simply by reading it. But each simile will also be accompanied by three sentences (from Twitter) to show you examples of how the simile can be used. If you are unsure which label to assign, you can read these sentences to get a better idea of how it is used by people on Twitter. You do not have to read the sentences if you are certain how to label a simile! If you do read the sentences, please keep in mind that these sentences are randomly selected and may not be typical. So label each simile based on YOUR OWN assessment of the simile, as well as the sample sentences. For example, suppose "Jane swims like a dolphin" is accompanied by 2 positive sentences and 1 negative sentence. Please choose the label that will be the most typical. It is ok if there are rare exceptions.

Category Definitions and Examples

The definitions for the 4 annotation labels are given below, with some examples. The words are shown in the root forms (i.e., looks → look, am/is/are → be). All personal pronouns (e.g, he, she, I) have been changed to "person".

1. POSITIVE: A positive sentiment is felt toward the subject of the simile majority of the times.

Examples:

(a) person swim like dolphin

(b) person look like model

(c) house be like castle

Explanation: Saying that a person swims like a dolphin or looks like a model is generally a compliment. Similarly, comparing someone's house to a castle is generally positive (e.g., implies that the house is nice, large, or comfortable).

2. NEGATIVE: A negative sentiment is felt toward the subject of the simile majority of the times.

   Examples:

   (a) room feel like sauna

   (b) person dress like hobo

   (c) hair look like bird's nest

   Explanation: Saying that a room feels like a sauna is usually a negative statement (e.g., the room is too hot or humid). Similarly, dressing like a hobo and having messy hair are generally negative comparisons.

3. NEITHER: Neither a clear positive nor a clear negative sentiment is felt toward the subject of the simile. Please note that this label should be used **very selectively**. Please assign either positive or negative label instead, if at all possible.

   Examples:

   (a) cloud look like turtle

   (b) weather smells like rain

   Explanation: Some similes will make comparisons that are not positive or negative, so these should be labeled NEITHER. For example, saying that a cloud looks like a turtle isn't a positive or negative statement. It is simply an observation. Similarly, "weather smell like rain" means that it may rain, which is not a positive or negative comparison, just an observation.

4. INVALID: The simile is malformed or the words do not represent a valid or meaningful comparison.

   Examples:

   (a) person feel like don't

   (b) person act like ima

   (c) friend don't like justin

   (d) person call like 5 times

   Explanation: some examples may not make sense, because they were identified automatically and our tools aren't perfect. If the word sequence isn't sensible (e.g., "person feel like don't", or "person act like ima"), label it as INVALID. Also, some may not be comparisons, so are not valid similes. For example, "like" is a verb in "friend don't like Justin" (it is not a comparison between "friend" and "Justin"). Similarly, "5 times" is the frequency of calling (there is no comparison between "person" and "5 times"). Minor difference between a word in a simile and the corresponding word in a tweet is okay, and that should NOT be the reason for invalid label. Check how the simile appears in the accompanying tweets. It may be easier to spot the invalid ones by checking the accompanying tweets. Remember that a valid simile will talk about the likeliness of two things, which will NOT be the case for invalid ones.

APPENDIX C

SIMILE IMPLICIT PROPERTY ANNOTATION GUIDELINES

Task Description and General Instructions

A simile is a figure of speech comparing two unlike things. For example, "this room feels like a sauna", or "he slept like a baby". You will be shown a list of 25 similes. Your task is to **give 2 properties** that describe possible meanings for each simile. For example, "this room feels like a sauna" could mean that the room is "hot" or that the room is "humid".

Each property should be a **single word**. Most properties will be adjectives (e.g., hot) or adverbs (e.g., loudly), but a property can also be a verb sometimes (e.g., freezing). Try to give 2 properties for most similes. But if you can only think of one, then enter "none" for the second property. If you can't think of any properties at all, then enter "none" in the first box, and leave the second box blank. (but please do this rarely).

All of the similes should be of the form "Noun + Verb + like + Noun". Each verb will be in its root form (e.g., "is" → "be", "feels" → "feel"), the pronouns (e.g., he, she) have been replaced with "person", and all words will be in lowercase. However, the similes were identified automatically and some similes may be malformed. If you find a simile where the first word or the last word is NOT a Noun, or the second word is NOT a Verb, then enter "invalid" in the first response box and leave the second box blank. For example, "room feel like kinda" and "happy look like baby" are both invalid similes.

Example Properties for Similes

Here are some example similes with appropriate responses:

(a) "room feel like sauna" → (1) hot, (2) humid

(b) "room feel like Antarctica" → (1) cold, (2) freezing

(c) "throat feel like sandpaper" → (1) rough, (2) scratchy

(d) "wallet feel like feather" → (1) light, (2) empty

(e) "person look like sister" → (1) none, (2)

(f) "today be like kinda" → (1) invalid, (2)

If you can think of more than 2 meanings for a simile, use your best judgement and select the two that seem the most likely meanings for the simile. Please try to give **specific** properties, rather than general terms. For example, for "throat feels like sandpaper", "rough" and "scratchy" are preferable to "bad". And do **not** give properties that reuse words from the simile. For example, for the simile "person sing like angel", do not give the property "angelic".

APPENDIX D

EMOTION TWEETS FROM HUMAN ANNOTATED DATA

D.1 Example Tweets Labeled with Affection

(a) @camspearman17 it's ok I'll fix it for you babe girlfriend #1 * I gotchu

(b) My team is forever the best family I will ever play with . #FBGM #weareoNe win or lose we are a family ! #sectionalbound

(c) Those goodnight text when I don't see him for the night <3 #cherishthelittletgings #husband

(d) Being left by your boyfriend so he can go watch football with your father <<<<<#thanks-babe

(e) she's taking a shot for me ... she doesn't drink so this is love >>>>> #myfavey #girlfriend

(f) I miss my baby I miss waking up to your face #boyfriend #muffin #babe #sleepovers #chinesefood #loveofmylife

(g) Cuddling on this rainy day with my baby ! #rainyday #cuddeling #boyfriend

(h) @Caliswag937 is my new #girlfriend #loveatfirstsite sorry @sam_bizzle

(i) #MentionSomeoneWhoCanAlwaysMakeYouSmile @Jed_Stoltzfus #obviously <3 #boyfriend

(j) #Mention10PeopleYouAreAfraidOfLosing My Babe , My Sweetheart , My Life , My Everything , My Girlfriend , My Wife @_MuchachaBonita #071012

D.2 Example Tweets Labeled with Anger/Rage

(a) May cannot come soon enough #sickofschool #graduation #soready

(b) That pic of that spider she just sent #FuckNo them damn things shouldn't be alive

(c) Sore throat and chest pain ... #damnit ... I cannot be sick at this moment ... #sigh ...

(d) @AndreaSturino : The stupid loud noise outside my bedroom window stopped as soon as I went to sleep in the spare bedroom ... #MagicTouch

(e) Working on no sleep someone just kill me #vamplife #nosleep

(f) erin andrews quit flirting with david freese thats my husband ! #backoff

(g) @nancyholtzman thank you ! We've been " dealing " with it by cosleeping , but I'd like my bed (and husband) back ! #newmoms

(h) Either divorce or leave me out of it ... I can't take this anymore !! #stopfighting #makelovenotwar

(i) Arrived home to find out my 75 year old grandparents are separating .. there goes 56 years of marriage . #WTF

(j) East side wild boyz we keep revolvers like cowboys betta run when ya hear the loud noise like Bruce Willis you can die hard #Flow

D.3 Example Tweets Labeled with Fear/Anxiety

(a) Headache back pain chest pain knee painear painshoulder pain #FallingApart

(b) @jaidestevenson me too ! im home alone haha #scared

(c) Should have started homework eariler ... gonna be up all night #nosleep #IBclasses

(d) Home alone all saturday and sunday #mumwise #needcompany #homealone #scaryshit

(e) #Nosleep #idontknow what to do

(f) Holy shit it cold out here !! #Witness #accident

(g) Have to change the channel every time a Sinister or Paranormal Activity preview comes on #homealone #helpmeimscared

(h) @apeavey101 @kristpeavey I was going to be home alone tonight so she said I could #scaryshit #cantdeal

(i) Came home to an empty house and I've been home alone for an hour and a half ... Maybe my family was murdered #worried #NotSureWhatToDo

(j) AHHHH THERE'S A HUGE SPIDER ON MY BED #WHYYYMEEE

D.4 Example Tweets Labeled with Joy

(a) well im home alone all weekend . its gunna be nice . #peaceful #silence

(b) Cap & gown ordered ! 8 more classes left . Holy cow I can't believe it's almost here ! #graduation #mediaspecialist

(c) Generic tweet that it's my favorite day of the week . #ACUWednesday #Armstrong #loveTheUniform #boyfriend #beatIowa

(d) it feels good af outside ... . #boyfriend .

(e) I don‘t wanna lose your love toniiiight ! #lyrics #SpaghettiWorks #Music #Great-Song

(f) So productive today ! #happy #dominationessay #studybreak #SUCCESS

(g) Is it too early to order my prom dress already ? #inLove

(h) @heathercrake @chrissieab8 @hrob92 loved this !!! Look what you did ya little jerk !!! #VFest #HomeAlone

(i) I have the prettiest dresses picked for the graduation dance #whichoneshouldichoose

(j) Bingo ! RT @KevinWGrossman As long as theres growth RT @DrJanice @TomBolt Trust = 2 simple words : Allow failure (or forgive failure) #tchat

D.5 Example Tweets Labeled with Sadness/Disappointment

(a) Dear @megamel88 can you stop outing my singleness , fear of roaches and drunkenness #mylifesucks #worst #husband #ever .

(b) when you make a twitter bc u got the best tweets , then you forget #failure #nolife

(c) and once again i am home alone . #dayfamilyprobs

(d) Looks like its me , the couch , and catters tonight . #homealone #foreverlonely #boyfriendcomehome #sadtweet

(e) #50ThingsAboutMyGirlfriend J'attends ces 50 choses sur moi !? Ah merde .. Pas de boyfriend ! #ForeverAlone #lol

(f) @stephyrose Aah that is bad aha I was gonna go for nail on the stairs in the foot ... #ouch #homealone

(g) Even with a divorce , you still have 2 kids together & you're always gonna have to be in eachothers lives . #getoverit #growupalready

(h) Bored , home alone .. nobody to talk to #Loner

(i) I swear I'm accident-prone .. lol , #colorguardprobz

(j) When your dad tells you he's leaving for a month to Cali <<<<<<<#daddysgirl #homealone #gayyyyy #iwannago

D.6 Example Tweets Labeled with "Other"

(a) " It's fine to celebrate #success but it is more important to heed the lessons of #failure . " — Bill Gates #quote

(b) Defo just gonna hire an accountant in the future ! #fail #exam #uni #resithereicome

(c) Thanks you @nicktrickey for allowing us to interview you . Video & transcript coming soon ! #automotive #news #interview

(d) #50ThingsAboutMyGirlfriend i don't love these hoes (prays that girlfriend does not see this tweet)

(e) #APrayer4Harrison Please get Harisson a job being a bouncer . Let him see his children grow . Let him wake & think of retirement #AMEN

(f) Home alone . Bumpin the Zelda soundtrack in my living room . About to draw on my notebook . #BringOnTheCats

(g) Busy day today ! Covers for 3 Annual Reports went out for client review and #corporate #web #video was edited and uploaded . #Success #Design

(h) Asking hubby to help due your hair = near divorce and a burnt scalp !! Goodbye red , hello chocolate brown #ouch

(i) Retirement used to mean a 401k , now we have 401k people on Obama's new retirement plan = welfare . #tcot #GOP

(j) @AskCheyB : A #ring is a #sign of a #covenant ! #wedding is a #celebration & #announcement covenant . A #marriage is ur #agreement 2 a covenant

D.7 Example Tweets Labeled with Multiple Emotions

Table D.1: Examples of tweets with multiple emotions.

Tweet	Emotion
Decided my obsession with marriage is actually an obsession with having a way out . Just kidding . I just hate it here . #sadtweet #fuckoff	Anger/Rage, Sadness/Disappointment
So if prom is in may , when's it okay to start looking for dresses ? Lol #anxious	Fear/Anxiety, Joy
Home alone , and i'm bored .. #Help #SomeoneComeHere #NoLife	Fear/Anxiety, Sadness/Disappointment
I texted my mom " MOM HELP " " HELP ME " " THEY'RE HERE " " WHAT AM I SUPPO " cause she left and i'm home alone . #Trolling	Anger/Rage, Joy
My 11:11 wish was made !!!! Praying it comes true #hoping #praying #wishing #loving #bestfriend #her #me #us #together #girlfriend	Affection, Joy
I think I just sent @Sweet_Tweet01 the cutest good-morning text ever ! (: #bestfriend #wife #loveeeeeher	Affection, Joy
they didn't called .]= #butthurt , #upset .. but i have another #interview tomorrow at 4 . so hope i do good on that , and they'll hire me . <3	Fear/Anxiety, Sadness/Disappointment
Rushing my sister to the chest pain er , was one of the scariest moments of my life #mysistermybestfriend	Affection, Fear/Anxiety
Lolling at my mom for saying I'm not allowed to go to wildwood after prom HAHA #jokesonyou	Anger/Rage, Joy
When your bestfriend brings you dinner because you've been home alone all day #shesthebest	Affection, Joy

APPENDIX E

TOP 100 LEARNED HASHTAG INDICATORS OF EMOTIONS

Table E.1: Top 100 learned emotion hashtags.

Affection	Anger/Rage	Fear/Anxiety	Joy	Sadness/Disapp.
#yourthebest	#godie	#hatespiders	#tripleblessed	#leftout
#notaprob	#donttalktome	#haunted	#tgfad	#foreverugly
#wishicouldbethere	#pieceofshit	#shittingmyself	#exicted	#singleprobs
#youdabest	#irritated	#worstfear	#thankful	#lonerlyfe
#otherhalf	#fuming	#scaresme	#24hours	#unloved
#youthebest	#hateliars	#nightmares	#birthdaycountdown	#jadeinbekasi
#bestfriendforever	#heated	#paranoid	#goodmood	#friendless
#flyhigh	#getoutofmylife	#hateneedles	#godisgood	#lonely
#loveyoulots	#angrytweet	#frightened	#greatmood	#teamlonely
#alwaysthere	#dontbothermewhen	#freakedout	#atlast	#heartbroken
#myotherhalf	#raging	#creepedout	#feelinggood	#notloved
#comehomesoon	#stupidbitch	#biggestfear	#happygirl	#singleprobz
#wuvyou	#madtweet	#sonervous	#godisgreat	#ineedfriends
#followher	#countkun	#shittingbricks	#lovemyfamily	#singleproblems
#alwaysandforever	#yourgross	#socreepy	#superhappy	#lonley
#alwayshere	#livid	#terrified	#newclothes	#needalife
#bestie	#screwyou	#waitinggame	#tentour	#lonertweet
#realfriend	#yousuck	#creeped	#newhair	#crushed
#missyousomuch	#badmood	#wimp	#liein	#miserable
#swimfast	#wankers	#nervous	#ecstatic	#letdown
#alwaysthereforme	#hateeverything	#freaked	#content	#singlepringle
#truebro	#somad	#scaredtodeath	#grateful	#catlady
#haveagoodone	#dickhead	#nerves	#12hours	#singlelife
#partnerincrime	#soannoyed	#mama	#prayertime	#loner
#loveyougirl	#backthefuckoff	#sleeplessnight	#happyboy	#alone
#gofollowher	#hadenough	#concerned	#glorytogod	#lovesucks
#bestbfever	#fuckedoff	#hiding	#gonnabegood	#singleforever
#bffl	#sopissed	#bigbaby	#superexcited	#ill
#loveyoualways	#anger	#freaky	#cantstopsmiling	#notthesame
#shesthebest	#fuckingstupid	#kindascared	#sunshine	#lonerforlife
#sheisthebest	#fuckyou	#soscared	#icanfeelit	#brokenhearted
#youdabomb	#hatepeople	#ghosts	#newhouse	#nofriends
#greatfriend	#knobs	#phobia	#churchflow	#hurting
#youramazing	#fucker	#strangefears	#lifesgood	#needfriends
#kisses	#grrrr	#sinister	#stressfree	#lifeisover
#comebacktome	#effoff	#panicmode	#clothes	#foreverlonely
#misshersomuch	#whatsyourproblem	#scaredshitless	#18th	#forgotten
#myangel	#prick	#shitmyself	#1day	#tomuchwork
#missyoualready	#grr	#scardycat	#birthday	#fucklove
#youreawesome	#fuckoff	#nosleepforme	#thankyoulord	#needcuddles
#bestsister	#doone	#imscared	#loveshopping	#sadlyf

Continued on next page.

Table E.1 Continued.

Affection	Anger/Rage	Fear/Anxiety	Joy	Sadness/Disapp.
#mwah	#soangry	#creepy	#siked	#loserstatus
#missu	#annoyedaf	#anxiety	#wohoo	#lowpoint
#cutie	#noonelikesyou	#nervouswreck	#can'twait	#cry
#twin	#pissoff	#areyoualive	#relaxed	#lifeofaloner
#thanksbaby	#wanker	#haunting2	#happyout	#brokenheart
#loveyoumore	#arseholes	#prayforme	#thankinggod	#goodfilm
#youreamazing	#asshole	#freakingout	#holidays	#memyselfandi
#yourethebest	#cunt	#curious	#excitedtweet	#teamloner
#loverher	#idontlikeyou	#scarymovies	#happytweet	#rejected
#needyou	#gofuckyourself	#ohno	#bestnewsever	#solonely
#proudofdemi	#fuckyourself	#notsafe	#familytime	#sadtimes
#mybabe	#dumbbitch	#baddriver	#thankfull	#lonerproblems
#loveandmissyou	#liar	#uhohh	#smiling	#teamnofriends
#bestfriendaward	#ihateyou	#hopingforthebest	#layin	#empty
#loveyoubabe	#physco	#shaking	#dancing	#lonerstatus
#foreverinmyheart	#pricks	#spiders	#tyj	#wahhh
#yougotthis	#ihatepeople	#shy	#missedthem	#fuckingshit
#mj23	#tfl	#scurred	#anniversary	#catladyproblems
#myrock	#youpissmeoff	#shitmypants	#lovemylife	#sadnight
#bigsis	#gotohell	#panic	#extremelyblessed	#dissapointing
#bestmom	#bitch	#scaredformylife	#woot	#hatebeingill
#loveyous	#donewithyou	#praying	#partyyy	#depressing
#mybitch	#getfucked	#wisdomteeth	#lovinglife	#allbymyself
#blessedgirlfriend	#dontlikeyou	#hatewaiting	#sun	#mysisterskeeper
#loveyouuuu	#cunts	#ombre	#yipee	#invisible
#lostwithoutyou	#foff	#notsleepingtonight	#sunsunsun	#betrayed
#followthem	#dick	#evildead	#sothankful	#tear
#bestfriend	#fuckeverything	#miniheartattack	#godfirst	#thatsajoke
#dontdie	#arghhh	#notready	#couldntbehappier	#novalentine
#bestboyfriendever	#2faced	#creepyshit	#greatday	#nolove
#missedher	#twofaced	#gonnadie	#cocktails	#nothingtodo
#hot30voteboost	#hateeveryone	#shittingit	#vaca	#loners
#lovehersomuch	#goaway	#strangefear	#perfectweekend	#bummed
#foreverandever	#twats	#scarymovie	#honored	#sadday
#hurryback	#gettofuck	#panicking	#gonnabegreat	#suicidal
#mybff	#diebitch	#ghost	#21st	#attached
#happyanniversary	#grumpy	#almostdied	#tgfl	#pale
#lover	#nitm	#anxietyattack	#tanning	#disapointed
#truefriend	#annoyed	#nosleeptonight	#drinks	#foreversingle
#bestfriendsforever	#aggravated	#hurryup	#letsdoitagain	#needacuddlebud
#bestgirlfriend	#youreannoying	#tooscared	#greatfriends	#noplans
#bestboyfriend	#justgoaway	#didntstudy	#weekendaway	#imsinglebecause
#lovehimsomuch	#shutthehellup	#scary	#newstart	#nosociallife
#bestgirlfriendever	#die	#thunderbuddy	#2days	#whyimsingle
#sisters	#shutup	#ah	#holiday	#unwanted
#awe	#ticked	#darkskies	#beautifulday	#nolife
#hesakeeperif	#imtryingtosleep	#scarystuff	#woop	#sadmoment
#bffz	#fuckingpissed	#druggedup	#yey	#thirdwheel
#sweet16	#stupidslut	#epicfight	#allsmiles	#ditched
#cousins	#dontmesswithme	#terrifying	#skiing	#paleproblems
#littlebrother	#shutyourmouth	#scaredycat	#yayy	#notfeelinggood
#happybirthdayed	#fakefriend	#cantstopthinking	#productiveday	#ruinedit
#pleasefollowme	#pissingmeoff	#politicalfilms	#newme	#teamnovalentine
#bestbrotherever	#yourabitch	#nameafear	#happysunday	#whatdoidonow
#myeverything	#roadrage	#screwed	#girlsnight	#mylifesucks
#missyoutoo	#leavemealone	#sketchy	#beach	#loserprobs

Continued on next page.

Table E.1 Continued.

Affection	Anger/Rage	Fear/Anxiety	Joy	Sadness/Disapp.
#missedyou	#thanksalot	#interested	#lovegod	#badnight
#2yearsunbroken	#obnoxious	#inpatient	#tan	#dateless
#bestbf	#wasteofspace	#tense	#radioactivetour	#privatepractice

Table E.2: Top 100 learned emotion hashtag patterns.

Affection	Anger/Rage	Fear/Anxiety	Joy	Sadness/Disapp.
you the best *	hate you *	nightmares *	ecstatic *	forever *
have a good one *	fuming *	paranoid *	feeling good *	unloved *
follow her *	dont talk *	scares *	at last *	heartbroken *
wish i could be *	angry *	freaked *	love my family *	te am lonely *
always and *	heated *	frightened *	happy girl *	nobody loves *
go follow her *	stupid bitch *	biggest fear *	good mood *	friendless *
real friend *	mad tweet *	creeped *	super excited *	broken heart *
true bro *	livid *	so creepy *	ten tour *	let down *
partner in *	so mad *	terrified *	super happy *	single life *
come back to me *	wankers *	spiders *	new hair *	good film *
always here *	so pissed *	so scared *	grateful *	need a life *
miss you so *	knobs *	kinda scared *	24 hours *	love sucks *
you da bomb *	you suck *	scariest thing *	thank you lord *	lonely *
great friend *	bad mood *	panic mode *	lie in *	disappoint *
always there *	grrrr *	nervous *	missed them *	loner *
lover her *	dickhead *	big baby *	happy boy *	no friends *
miss her so *	hate people *	pray for me *	honored *	need friends *
cutie *	piss off *	freaking out *	i can feel *	fuck love *
youre the best *	pricks *	anxiety *	new clothes *	nothing to do *
birthday shout *	beyond pissed *	nerves *	family time *	wahhh *
miss you already *	leave me alone *	scardy *	can't wait *	miserable *
love and miss you *	fucker *	concerned *	content *	lonley *
you got this *	prick *	scurred *	happy tweet *	sad day *
bestie *	had enough *	haunted *	stress free *	worst feelings *
fly high *	bad morning *	sinister *	thanking *	hate valentines *
love yous *	wanker *	insidious *	relaxed *	low point *
true friend *	i dont like you *	ghosts *	best news *	hurting *
youre awesome *	youre annoying *	shitting it *	holidays *	rejected *
my bitch *	fuck you bitch *	no sleep tonight *	love my life *	sad times *
see you soon *	do one *	creepy as *	loving life *	alone *
shes the *	aggravated *	hoping for the *	happy out *	no life *
my angel *	raging *	oh no *	12 hours *	bummed *
luhh you *	you piss *	old house *	lay in *	disapointed *
best friends for *	no one likes you *	curious *	so thankful *	no valentine *
best girlfriend *	go fuck yourself *	im scared *	love shopping *	normal day *
miss you too *	annoyed *	no sleep for *	gonna be great *	no plans *
follow them *	fumin *	phobia *	great day *	broken hearted *
follow this *	dumb bitch *	my car *	spoiling *	dissapointing *
love him to *	sick of it *	bad driver *	give thanks *	single status *
sisters for *	frustrated *	hate waiting *	sun sun *	tomorrows a *
best cousin *	dont bother me *	going to fail *	glory to *	neglected *
best bf *	fuck off *	too scared *	god first *	loners *
best friend *	twats *	mini heart *	1 day *	used to it *
missed her *	hate everyone *	sick to my *	new start *	wanna cry *
missed you *	ill kill *	hurrry up *	great friends *	ruined it *

Continued on next page.

Table E.2 Continued.

Affection	Anger/Rage	Fear/Anxiety	Joy	Sadness/Disapp.
little brother *	pissing me *	thunder buddy *	productive day *	im single *
best boyfriend *	shut the fuck *	interested *	happy sunday *	sadness *
sweet 16 *	arghhh *	antsy *	feels good *	my sisters *
gotta love her *	ruin my *	scaredy *	dnow 2013 *	depressing *
she da best *	you fucking *	haunting *	busy week *	so lonely *
reunite *	no one gives *	shitting *	jello shots *	sad moment *
lets hang *	late for work *	gonna die *	woohoo *	my life sucks *
adorable *	idiots *	almost died *	sunshine *	private practice *
hugs and *	asshole *	givin *	lifes great *	third wheel *
daddys little *	not in the mood *	didnt study *	relieved *	betrayed *
best valentines *	i hate people *	so stressed *	rested *	single 4 *
bros for *	fake people *	car sick *	perfect sunday *	no social life *
cake cake cake *	road rage *	not sleeping *	beautiful day *	unwanted *
hurry back *	dont piss me off *	slender *	organised *	ditched *
get better soon *	cunts *	wisdom tooth *	new home *	heartbreaking *
miss you *	just go away *	needle *	yaaaay *	hard times *
miss youuu *	thanks alot *	hair dye *	accomplished *	dreams of a *
love him so *	so done *	please help me *	so happy *	what happend *
sisters *	shut up *	praying *	new me *	bad night *
kisses *	two faced *	bricking *	love life *	need someone *
best friends *	no one cares *	scary shit *	wooooo *	cat lover *
separation *	so annoying *	not prepared *	clothes *	phoneless *
growing up so *	get to fuck *	scary movies *	thank you jesus *	emosh *
they grow up so *	tryin to *	wisdom teeth *	about time *	bored out *
brother sister *	youre not cool *	crossing my *	happy me *	lifeless *
youre perfect *	annoys *	ready to play *	lucky girl *	ignored *
besties *	go the fuck *	restless *	been waiting *	i wanted *
need you *	obnoxious *	creepy *	love days *	single *
you rock *	just stop *	cant study *	all smiles *	gloomy *
bestfriends *	i hate you *	inpatient *	smiling *	not fun *
love ya *	shut the hell *	scary *	couldnt ask *	team single *
someone i *	trying to sleep *	never sleeping *	happy dance *	self pity *
inseparable *	go away *	paranoia *	praise the *	cat lady *
thank you so much *	pet peeve *	drugged *	very excited *	no friend *
the sweetest *	grow the *	worst nightmare *	happy kid *	dateless *
twinning *	you stink *	uh oh *	anniversary *	ima loser *
channing tatum *	2 faced *	holy crap *	love god *	too sad *
long lost *	bitch mode *	scarred for *	energetic *	ridin solo *
seeya *	douche *	cant take this *	i can get *	cold and *
miss her *	trying to study *	too windy *	so grateful *	bawling *
cousins *	biggest pet *	dark skies *	thankyou god *	not a good day *
harry 1 *	fuck everything *	terrifying *	sleeping in *	sad life *
ride or die *	so loud *	whats going *	3day weekend *	dont need a *
ma homie *	just shut up *	hiding *	no stress *	no fun *
were cool *	morons *	say a *	missed him *	guess not *
change that *	get a life *	not normal *	satisfied *	oh fucking *
only the best *	indirect tweet *	brace face *	exited *	entertain me *
best sister *	its annoying *	sketchy *	2 days *	i give up *
real love *	shame shame *	unprepared *	put god *	gutting *
miss u *	calm the fuck *	too close for *	so lucky *	bored af *
my girl *	your stupid *	traumatized *	9 days *	life sucks *
please follow me *	ticked *	too much on *	amazed *	unmotivated *
heart to *	burn in hell *	evil dead *	stocked *	what happened *

Continued on next page.

Table E.2 Continued.

Affection	Anger/Rage	Fear/Anxiety	Joy	Sadness/Disapp.
love her *	i will kill *	day after *	i cant wait *	forgotten *
hes perfect *	one thing i hate *	not feeling well *	new friends *	nothing new *

APPENDIX F

EXAMPLES OF SIMILES ANNOTATED WITH HUMAN ANNOTATORS FOR AFFECTIVE POLARITY

Table F.1: Examples of similes labeled with gold labels for positive/negative affective polarity.

Positive	**Negative**
<PERSON, get, model>	<PERSON, look, hooker>
<football, be, nfl>	<PERSON, feel, burden>
<PERSON, wear, battle wound>	<PERSON, feel, bear>
<PERSON, feel, superman>	<IT, be, nothing>
<PERSON, be, popeye>	<PERSON, feel, such a dumbass>
<PERSON, be, family>	<PERSON, look, kid>
<PERSON, get, jesus>	<PERSON, feel, walk zombie>
<degree, feel, summer>	<puberty, hit, truck>
<PERSON, be, kanye>	<IT, look, fish>
<PERSON, be, barbie>	<IT, smell, pee>
<love, be, flower>	<PERSON, look, bratz doll>
<fear, be, love>	<IT, feel, lifetime>
<IT, be, lot of relationship>	<PERSON, be, fire and gasoline>
<PERSON, float, feather>	<PERSON, be, rest>
<PERSON, look, mila kunis>	<PERSON, shock, electric eel>
<PERSON, look, leonardo dicaprio>	<PERSON, look, serial killer>
<bed, be, cloud>	<PERSON, feel, outsider>
<PERSON, love, sister>	<PERSON, look, loaf of bread>
<PERSON, be, pea and carrot>	<PERSON, fight, married couple>
<IT, smell, spring>	<IT, sound, problem>
<PERSON, be, boss>	<IT, be, hunger games>
<kobe, be, lebron>	<IT, be, ghost town>
<IT, feel, weekend>	<PERSON, look, alien>
<IT, be, game>	<PERSON, look, poo>
<PERSON, take, champ>	<PERSON, feel, such a nerd>
<player, be, club>	<PERSON, sound, goat>
<PERSON, feel, virgin>	<PERSON, look, mess>
<PERSON, be, mj>	<people, be, slinky>
<nirvana, smell, teen spirit>	<PERSON, love, nobody>
<IT, be, concert>	<PERSON, act, brat>

Continued on next page.

Table F.1 Continued.

Positive	Negative
<IT, be, winter wonderland>	<PERSON, look, pug>
<PERSON, feel, bird>	<PERSON, look, mouse>
<PERSON, tweet, budgie>	<PERSON, look, chucky>
<PERSON, know, back of she hand>	<PERSON, be, little girl>
<PERSON, be, older brother>	<PERSON, look, casper>
<PERSON, look, jesus>	<PERSON, smell, wet dog>
<PERSON, be, sunshine>	<IT, look, trash>
<IT, look, snow>	<PERSON, feel, nobody>
<IT, be, college>	<IT, be, work>
<tomorrow, sound, plan>	<PERSON, look, idiot>
<PERSON, feel, champion>	<PERSON, sound, nerd>
<PERSON, be, fan>	<PERSON, treat, joke>
<PERSON, be, snowflake>	<PERSON, feel, puke>
<PERSON, look, bag of money>	<PERSON, feel, prisoner>
<PERSON, be, mom>	<money, play, dummy>
<happiness, hit, train>	<IT, do not feel, home>
<love, be, pride>	<PERSON, look, raccoon>
<PERSON, look, jennifer aniston>	<PERSON, feel, peasant>
<pen, be, sword>	<PERSON, change, girl change clothes>
<wisdom, be, silver or gold>	<PERSON, act, girl>

Table F.2: Examples of similes labeled with neutral/invalid.

Neutral	Invalid
<ambition, be, bird>	<PERSON, be, oooo>
<IT, look, bird>	<PERSON, be, all of we>
<PERSON, think, oomf>	<PERSON, feel, cryin>
<PERSON, be, guy>	<PERSON, be, half>
<twitter, be, fridge>	<happiness, be, you own>
<PERSON, look, you brother>	<christmas, be, month>
<PERSON, feel, guy>	<IT, be, fr>
<IT, be, cross>	<PERSON, be, ummm>
<PERSON, feel, lil girl>	<PERSON, feel, trip>
<PERSON, be, shoe>	<spirit, be, proud>
<PERSON, be, domino>	<success, should be, you fear of failure>
<PERSON, look, cross>	<conversation, smile, idiot>
<IT, look, duck>	<PERSON, feel, everyone>
<house, smell, pizza>	<PERSON, feel, ballin>
<PERSON, look, chinese>	<PERSON, do, month>
<mind, be, parachute>	<PERSON, be, suck>
<PERSON, look, chick>	<PERSON, be, whattt>
<PERSON, feel, cat>	<PERSON, be, wut>

Continued on next page.

Table F.2 Continued.

Neutral	Invalid
<PERSON, seem, kind of person>	<PERSON, be, financial aid>
<PERSON, feel, some people>	<PERSON, have, last year>
<PERSON, look, cat>	<IT, be, 5am>
<IT, slide, sunroof>	<PERSON, addict, addict>
<PERSON, look, different person>	<PERSON, make, last month>
<IT, look, face>	<PERSON, be, nahhhhhh>
<IT, be, fact>	<IT, be, tho>
<PERSON, feel, different person>	<PERSON, be, omfg>
<instagram, be, snapchat>	<PERSON, be, dancing>
<PERSON, be, number>	<PERSON, sound, youre>
<IT, be, competition>	<PERSON, make, year>
<PERSON, look, girl>	<yes, sound, plan>
<IT, sound, job>	<one, be, last>
<PERSON, be, texas>	<PERSON, be, last>
<PERSON, be, tweetuser tweetuser>	<school, get, thishttp>
<IT, be, time>	<PERSON, be, nope>
<IT, look, harry>	<PERSON, be, umm>
<nothing, be, guy>	<IT, be, do>
<PERSON, look, boosie>	<PERSON, be, dayum>
<PERSON, look, each other>	<IT, be, ok>
<PERSON, look, jade>	<PERSON, be, ummmm>
<PERSON, be, sophomore>	<PERSON, feel, fact>
<question, be, answer>	<PERSON, feel, alot of people>
<PERSON, be, sam>	<PERSON, be, tht>
<life, be, boat>	<PERSON, look, youre>
<PERSON, look, lil girl>	<IT, be, chill>
<PERSON, look, #oomf>	<PERSON, be, uhhhhhh>
<PERSON, feel, dr>	<PERSON, look, all the time>
<PERSON, look, dora>	<PERSON, have, idk>
<PERSON, look, jake>	<PERSON, be, smh>
<PERSON, act, somebody>	<PERSON, see, week>
<IT, be, live>	<IT, 's be, hour>

APPENDIX G

EXAMPLES OF SIMILES ANNOTATED WITH IMPLICIT PROPERTIES FROM THE HUMAN ANNOTATED DATA

Table G.1: Examples of similes and their implicit properties from human judgements.

Simile	**Implicit Properties**
<laugh, be, music>	melodic, pleasant, pleasing, harmonious, silvery, dulcet, beautiful, tinkly, enjoyable, melodious
<shower, feel, heaven>	ideal, soothing, apt, rejuvenating, pleasant, relaxing, wonderful, great, warm, blissful, glorious, gentle
<person, sound, prophet>	insightful, informative, teaching, foreseeing, wise, knowing, prescient, divinitory, sage, enlightened, philisophical, knowledgable
<fan, look, idiot>	stupid, silly, dumb, overexcited, ridiculous, fool, unattractive, moron
<PERSON, dress, hooker>	trashily, scantily, sexy, trashy, flamboyant, flashy, risqué, cheap, slinky, unprofessional, provocatively, slutty, suggestively
<PERSON, be, shark>	primordial, scheming, dangerous, cold, stealthy, toothy, opportunistic, greedy, aggressive, hunting, sneaky, bloodthirsty, predatory
<PERSON, look, mr>	formal, handsome, fancy, responsible, serious, manly, man, gentleman, male, professional
<PERSON, feel, chocolate>	silky, bitter, velvety, indulgent, warm, dark, hard, sweet, soft, smooth
<PERSON, look, puke>	nauseating, green, gross, ugly, messy, sickly, disgusting, unattractive
<PERSON, feel, child>	young, carefree, innocent, helpless, youthful, patronized, immature, dependent, naive
<house, smell, dog>	repulsive, smelly, stinky, musty, wet, disgustive, earthy, moldy, damp, musky, dank
<PERSON, wait, hour>	patient, forever, unending, long

Continued on next page.

Table G.1 Continued.

Simile	Implicit Properties
<baseball, be, church>	applauded, religious, traditional, holy, customary, sacred, worshipped
<PERSON, look, nerd>	dorky, uncool, silly, informative, dork, intellectual, smart, intelligent, bookish, unfashionable, bookworm, geek, antisocial
<people, smell, dog>	dirty, repulsive, smelly, gross, musty, wet, moldy, sour, dank, disgusting, moist
<PERSON, be, mama>	comforting, kind, loving, warm, nurturing, welcoming, protective, controlling, maternal
<hair, smell, fire>	smoky, hot, ashy, pungent, smokey, burnt
<PERSON, look, potato>	lumpy, obese, round, blobby, plain, fat, dumpy, dull
<pee, smell, cheerios>	natural, yeasty, oaty, distinctive, nutty, cardboard, sweet
<PERSON, look, alien>	green, otherworldly, ugly, martian, strange, foreign, weird, freak, unusual
<niggas, drop, fly>	fast, routinely, frequently, quickly, easily, dead, hard, lightweight, weak, gone
<niggas, be, squad>	gang, united, cohesive, family, together, faithful, crew, cooperative
<PERSON, sound, fun>	comical, entertaining, friendly, hilarious, happy, joyous, lively, enjoyable, bubbly, amusing, playful
<night, be, movie>	idealized, unbelievable, silent, unrealistic, adventurous, long, eventful, dark, fantastic, fake, unreal
<life, be, circle>	continous, endless, unending, round, continuous, repeating, complete, repetitive, cyclic, smooth
<PERSON, feel, friend>	supportive, familiar, recognizable, loving, warm, caring, close, cordial
<PERSON, smell, bullshit>	insincere, smelly, poopy, repugnant, false, untruthful, foul, disgusting, fake, liar, nasty, lying
<PERSON, dance, diva>	egotistic, proud, excellent, sexy, confidently, dramatic, competent, magnificent, good, accomplished, brazenly, athletic, smooth
<PERSON, be, elf>	small, slight, otherworldly, ugly, mythical, tiny, short, demure, petite
<friend, be, condom>	secure, tight, snug, safe, shielding, protective, wall, safeguard
<PERSON, be, vampire>	enthralling, silent, hypnotic, dangerous, spooky, toothy, dark, menacing, evil, pale, bloodthirsty

Continued on next page.

Table G.1 Continued.

Simile	Implicit Properties
<time, be, river>	unending, flowing, fast, winding, rushing, moving, swift
<today, feel, weekend>	carefree, relaxing, relaxed, unhurried, calm, short, free, fun
<mind, be, religion>	supportive, obsessed, preoccupied, powerful, strength, worshiped, mysterious, vast, believing, peaceful, sacred
<voice, be, music>	soothing, melodic, lilting, good, pleasing, beautiful, enjoyable, sonorous, melodious, soft, joyful
<PERSON, smell, ashtray>	repulsive, smoky, bad, musty, reeking, bitter, acrid, stale, smokey
<praise, be, sunlight>	energizing, rejuvenating, bright, cheerful, pleasing, warm, fulfilling, brightening
<PERSON, look, cyclops>	monstrous, one-eyed, ugly, dim, unattractive, humungous, hideous, deformed
<boy, be, commercial>	loud, knockoff, ubiquitous, demonstrates, exaggerated, short, fake, selling, pushy, annoying
<PERSON, look, sin>	beaten, dangerous, awful, ugly, unholy, seductive, enticing, mean, evil, scary, tempting, terrible
<PERSON, feel, hitler>	egotistic, cruel, murderer, unstable, vindictive, manic, evil, inhuman, strict, opressive, terrible
<PERSON, be, shoot>	loud, explosive, new, sniper, exact, fire, tall, fresh, thin, sharp
<PERSON, be, bit>	small, constraining, compact, insignificant, tiny, leading, little
<toe, look, bean>	small, irregular, brown, cute, long, round, slender, tiny, dark, short, little
<PERSON, smell, apple>	tangy, aromatic, sour, fruity, juicy, tart, sweet, fresh
<car, smell, coconut>	fruity, sharp, tropical, creamy, oily, sweet, fresh
<PERSON, feel, haha>	amused, happy, content
<body, stick, glue>	suction, spittle, tight, holding, fast, clingy, tacky, securely, strongly, tightly, adhesive, close, needy
<PERSON, look, sir>	prestigious, masculine, polite, formal, lord, proper, manly, gentleman, male, classy, gentlemanly
<PERSON, feel, peasant>	lowly, simple, weak, servile, hungry, low, poor, destitute, useless

REFERENCES

Addison, C. (1993). From literal to figurative: An introduction to the study of simile. *College English 55*(4), 402–419.

Adreevskaia, A. and S. Bergler (2006). Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 209–216.

Akritidis, L., D. Katsaros, and P. Bozanis (2011). Effective rank aggregation for metasearching. *Journal of Systems and Software 84*(1), 130–143.

Albrecht, S. L. et al. (2010). *Handbook of employee engagement: Perspectives, issues, research and practice*. Edward Elgar Publishing.

Baccianella, S., A. Esuli, and F. Sebastiani (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Volume 10, pp. 2200–2204.

Banea, C., R. Mihalcea, and J. Wiebe (2008, May). A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08)*, Marrakech, Morocco. European Language Resources Association (ELRA). ACL Anthology Identifier: L08-1086.

Barbosa, L. and J. Feng (2010). Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, Stroudsburg, PA, USA, pp. 36–44. Association for Computational Linguistics.

Beardsley, M. C. (1981). *Aesthetics, problems in the philosophy of criticism*. Hackett Publishing.

Berg-Kirkpatrick, T., D. Burkett, and D. Klein (2012). An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 995–1005. Association for Computational Linguistics.

Bollen, J., H. Mao, and A. Pepe (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the International AAAI Conference on Web and Social Media*, pp. 450–453.

Bredin, H. (1998). Comparisons and similes. *Lingua 105*(1), 67–78.

Brody, S. and N. Elhadad (2010). An unsupervised aspect-sentiment model for online reviews. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 804–812. Association for Computational Linguistics.

Brun, C., D. N. Popa, and C. Roux (2014). Xrce: Hybrid classification for aspect-based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 838–842.

Cambria, E., J. Fu, F. Bisio, and S. Poria (2015). Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 508–514.

Cambria, E. and A. Hussain (2012). *Sentic computing: Techniques, tools, and applications*, Volume 2. Springer Science & Business Media.

Cambria, E., R. Speer, C. Havasi, and A. Hussain (2010). Senticnet: A publicly available semantic resource for opinion mining. In *Proceedings of the AAAI fall symposium: commonsense knowledge*.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics 22*(2), 249–254.

Carlson, A., J. Betteridge, E. R. Hruschka, Jr., and T. M. Mitchell (2009). Coupling semi-supervised learning of categories and relations. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, SemiSupLearn '09, Stroudsburg, PA, USA, pp. 1–9. Association for Computational Linguistics.

Carlson, A., J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell (2010). Toward an architecture for never-ending language learning. In *Proceedings of AAAI conference on artificial intelligence*, Volume 5, pp. 3.

Carter, S., W. Weerkamp, and M. Tsagkias (2013). Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation 47*(1), 195–215.

Castellucci, G., S. Filice, D. Croce, and R. Basili (2014). Unitor: Aspect based sentiment analysis with structured learning. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 761–767.

Chang, C.-C. and C.-J. Lin (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST) 2*(3), 27.

Chiappe, D. L. and J. M. Kennedy (2000). Are metaphors elliptical similes? *Journal of Psycholinguistic Research 29*(4), 371–398.

Chikersal, P., S. Poria, E. Cambria, A. Gelbukh, and C. E. Siong (2015). Modelling public sentiment in twitter: using linguistic patterns to enhance supervised learning. In *Computational Linguistics and Intelligent Text Processing*, pp. 49–65. Springer.

Chklovski, T. (2003). Learner: a system for acquiring commonsense knowledge by analogy. In *Proceedings of the 2nd international conference on Knowledge capture*, pp. 4–12. ACM.

Choi, Y., L. Deng, and J. Wiebe (2014). Lexical acquisition for opinion inference: A sense-level lexicon of benefactive and malefactive events. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, pp. 107–112.

Choi, Y. and J. Wiebe (2014, October). +/-effectwordnet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1181–1191. Association for Computational Linguistics.

Cui, H., V. Mittal, and M. Datar (2006). Comparative experiments on sentiment classification for online product reviews. In *Proceedings of the AAAI conference on artificial intelligence*, Volume 6, pp. 1265–1270.

Davidov, D., O. Tsur, and A. Rappoport (2010). Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 241–249. Association for Computational Linguistics.

Deng, L. and J. Wiebe (2015, September). Joint prediction for entity/event-level sentiment analysis using probabilistic soft logic models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp. 179–189. Association for Computational Linguistics.

Desmet, B. and V. Hoste (2013). Emotion detection in suicide notes. *Expert Systems with Applications 40*(16), 6351–6358.

Ding, H. and E. Riloff (2016). Acquiring knowledge of affective events from blogs using label propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Dwork, C., R. Kumar, M. Naor, and D. Sivakumar (2001). Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pp. 613–622. ACM.

Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion 6*(3-4), 169–200.

Esmin, A. A. A., R. de Oliveira, and S. Matwin (2012). Hierarchical classification approach to emotion recognition in twitter. In *Proceedings of Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, Volume 2, pp. 381–385. IEEE.

Etzioni, O., M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates (2004). Web-scale information extraction in knowitall:(preliminary results). In *Proceedings of the 13th international conference on World Wide Web*, pp. 100–110. ACM.

Etzioni, O., M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence 165*(1), 91–134.

Fahrni, A. and M. Klenner (2008). Old wine or warm beer: Target-specific sentiment analysis of adjectives. In *Proceedings of the Symposium on Affective Language in Human and Machine, AISB*, pp. 60–63.

Fan, R.-E., K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin (2008). Liblinear: A library for large linear classification. *The Journal of Machine Learning Research 9*, 1871–1874.

Feng, S., J. S. Kang, P. Kuznetsova, and Y. Choi (2013, August). Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Sofia, Bulgaria, pp. 1774–1784. Association for Computational Linguistics.

Fishelov, D. (2007). Shall i compare thee? simile understanding and semantic categories. *Journal of literary semantics 36*(1), 71–87.

Fraisse, A. and P. Paroubek (2014). Twitter as a comparable corpus to build multilingual affective lexicons. In *Proceedings of The 7th Workshop on Building and Using Comparable Corpora*, pp. 26–31.

Fukuhara, T., H. Nakagawa, and T. Nishida (2007). Understanding sentiment of people from news articles: Temporal sentiment analysis of social events. In *PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON WEBLOGS AND SOCIAL MEDIA*.

Gagné, C. L. (2002). Metaphoric interpretations of comparison-based combinations. *Metaphor and Symbol 17*(3), 161–178.

Ghosh, A., G. Li, T. Veale, P. Rosso, E. Shutova, J. Barnden, and A. Reyes (2015, June). Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, Colorado, pp. 470–478. Association for Computational Linguistics.

Glucksberg, S., M. S. McGlone, and D. Manfredi (1997). Property attribution in metaphor comprehension. *Journal of memory and language 36*(1), 50–67.

Go, A., R. Bhayani, and L. Huang (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford 1*, 12.

Godbole, N., M. Srinivasaiah, and S. Skiena (2007). Large-scale sentiment analysis for news and blogs. *ICWSM 7*, 21.

Goyal, A., E. Riloff, H. Daume III, and N. Gilbert (2010). Toward plot units: Automatic affect state analysis. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pp. 17–25. Association for Computational Linguistics.

Hanks, P. (2005). Similes and sets: The english preposition 'like'. *Languages and Linguistics: Festschrift for Fr. Cermak. Charles University, Prague*.

Hu, M. and B. Liu (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177. ACM.

Israel, M., J. R. Harding, and V. Tobin (2004). On simile. *Language, culture, and mind 100*, 123–35.

Jiang, L., M. Yu, M. Zhou, X. Liu, and T. Zhao (2011). Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 151–160. Association for Computational Linguistics.

Jijkoun, V., M. de Rijke, and W. Weerkamp (2010). Generating focused topic-specific sentiment lexicons. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, Stroudsburg, PA, USA, pp. 585–594. Association for Computational Linguistics.

Jo, Y. and A. H. Oh (2011). Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 815–824. ACM.

Kanayama, H. and T. Nasukawa (2006). Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, Stroudsburg, PA, USA, pp. 355–363. Association for Computational Linguistics.

Kim, S.-M. and E. Hovy (2004). Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, pp. 1367. Association for Computational Linguistics.

Kiritchenko, S., X. Zhu, C. Cherry, and S. Mohammad (2014). Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 437–442. Association for Computational Linguistics and Dublin City University Dublin, Ireland.

Klementiev, A., D. Roth, and K. Small (2008). Unsupervised rank aggregation with distance-based models. In *Proceedings of the 25th international conference on Machine learning*, pp. 472–479. ACM.

Kouloumpis, E., T. Wilson, and J. Moore (2011). Twitter sentiment analysis: The good the bad and the omg! *ICWSM 11*, 538–541.

Kozareva, Z., B. Navarro, S. Vázquez, and A. Montoyo (2007). Ua-zbsa: a headline emotion classification through web information. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 334–337. Association for Computational Linguistics.

Kozareva, Z., E. Riloff, and E. Hovy (2008, June). Semantic class learning from the web with hyponym pattern linkage graphs. In *Proceedings of ACL-08: HLT*, Columbus, Ohio, pp. 1048–1056. Association for Computational Linguistics.

Lebanon, G. and J. Lafferty (2002). Cranking: Combining rankings using conditional probability models on permutations. In *Proceedings of the Nineteenth International Conference on Machine Learning (ICML-2002)*, Volume 2, pp. 363–370.

Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, et al. (2015). Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web 6*(2), 167–195.

Levy, O. and Y. Goldberg (2014). Dependencybased word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Volume 2, pp. 302–308.

Levy, O., Y. Goldberg, and I. Dagan (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics 3*, 211–225.

Li, B., H. Kuang, Y. Zhang, J. Chen, and X. Tang (2012). Using similes to extract basic sentiments across languages. In *Web Information Systems and Mining*, pp. 536–542. Springer.

Li, J., A. Ritter, C. Cardie, and E. Hovy (2014). Major life event extraction from twitter based on congratulations/condolences speech acts. In *Proceedings of Empirical Methods in Natural Language Processing*.

Lin, K. H.-Y., C. Yang, and H.-H. Chen (2008). Emotion classification of online news articles from the reader's perspective. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pp. 220–226. IEEE Computer Society.

Liu, B. and L. Zhang (2012). A survey of opinion mining and sentiment analysis. In *Mining text data*, pp. 415–463. Springer.

Liu, H. and P. Singh (2004). Conceptneta practical commonsense reasoning tool-kit. *BT technology journal 22*(4), 211–226.

Liu, Y.-T., T.-Y. Liu, T. Qin, Z.-M. Ma, and H. Li (2007). Supervised rank aggregation. In *Proceedings of the 16th international conference on World Wide Web*, pp. 481–490. ACM.

Lu, Y., M. Castellanos, U. Dayal, and C. Zhai (2011). Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th international conference on World wide web*, pp. 347–356. ACM.

Mahdisoltani, F., J. Biega, and F. Suchanek (2014). Yago3: A knowledge base from multilingual wikipedias. In *Proceedings of the 7th Biennial Conference on Innovative Data Systems Research*. CIDR 2015.

Maks, I. and P. Vossen (2013). Sentiment analysis of reviews: Should we analyze writer intentions or reader perceptions? In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pp. 415–419. INCOMA Ltd. Shoumen, BULGARIA.

Manning, C. D., M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60.

McIntosh, T. (2010). Unsupervised discovery of negative categories in lexicon bootstrapping. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, Stroudsburg, PA, USA, pp. 356–365. Association for Computational Linguistics.

McIntosh, T. and J. R. Curran (2008, December). Weighted mutual exclusion bootstrapping for domain independent lexicon and template acquisition. In *Proceedings of the Australasian Language Technology Association Workshop 2008*, Hobart, Australia, pp. 97–105.

McIntosh, T. and J. R. Curran (2009). Reducing semantic drift with bagging and distributional similarity. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language*

Processing of the AFNLP: Volume 1 - Volume 1, ACL '09, Stroudsburg, PA, USA, pp. 396–404. Association for Computational Linguistics.

Medhat, W., A. Hassan, and H. Korashy (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal 5*(4), 1093–1113.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM 38*(11), 39–41.

Mitchell, J. and M. Lapata (2008). Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pp. 236–244. Association for Computational Linguistics.

Mohammad, S. (2012a, 7-8 June). #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, Montréal, Canada, pp. 246–255. Association for Computational Linguistics.

Mohammad, S. (2012b). Portable features for classifying emotional text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Mohammad, S., C. Dunne, and B. Dorr (2009). Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 599–608. Association for Computational Linguistics.

Mohammad, S. M. (2011). Even the abstract have colour: Consensus inword–colour associations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: shortpapers*, pp. 368–373.

Mohammad, S. M., S. Kiritchenko, and X. Zhu (2013, June). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA.

Mohammad, S. M. and P. D. Turney (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence 29*(3), 436–465.

Mohammad, S. M. and T. W. Yang (2011). Tracking sentiment in mail: how genders differ on emotional axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (ACL-HLT 2011*, pp. 70–79.

Murphy, T. and J. R. Curran (2007). Experiments in mutual exclusion bootstrapping. In *Proceedings of the Australasian Language Technology Workshop (ALTW)*, pp. 66–74.

Na, J.-C., C. Khoo, and P. H. J. Wu (2005). Use of negation phrases in automatic sentiment classification of product reviews. *Library Collections, Acquisitions, and Technical Services 29*(2), 180–191.

Nakov, P., S. Rosenthal, Z. Kozareva, V. Stoyanov, A. Ritter, and T. Wilson (2013, June). Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, USA, pp. 312–320. Association for Computational Linguistics.

Neviarouskaya, A., H. Prendinger, and M. Ishizuka (2011). Sentiful: A lexicon for sentiment analysis. *Affective Computing, IEEE Transactions on 2*(1), 22–36.

Niculae, V. (2013). Comparison pattern matching and creative simile recognition. In *Proceedings of the Joint Symposium on Semantic Processing.*

Niculae, V. and C. Danescu-Niculescu-Mizil (2014). Brighter than gold: Figurative language in user generated comparisons. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2008–2018. Association for Computational Linguistics.

Niculae, V. and V. Yaneva (2013). Computational considerations of comparisons and similes. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pp. 89–95. Association for Computational Linguistics.

Nie, J.-Y., J. Gao, J. Zhang, and M. Zhou (2000). On the use of words and n-grams for chinese information retrieval. In *Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, pp. 141–148. ACM.

Nielsen, F. Å. (2011). "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs",. In *Proceedings of ESWC2011 Workshop on Making Sense of Microposts: Big things come in small packages.*

Ohana, B. and B. Tierney (2009). Sentiment classification of reviews using sentiwordnet. In *Proceedings of the 9th. IT & T Conference*, pp. 13.

Ortony, A. (1993). *Metaphor and thought.* Cambridge University Press.

Owoputi, O., B. OConnor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-2013).*

ODonoghue, J. (2009). Is a metaphor (like) a simile? differences in meaning, effects and processing. *UCL Working Papers in Linguistics 21*, 125–149.

Pak, A. and P. Paroubek (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Volume 10, pp. 1320–1326.

Pang, B. and L. Lee (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval 2*(1-2), 1–135.

Parrott, W. G. (2001). *Emotions in social psychology: Essential readings.* Psychology Press.

Pasca, M. (2004). Acquisition of categorized named entities for web search. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, New York, NY, USA, pp. 137–145. ACM.

Paul, A. M. (1970). Figurative language. *Philosophy & Rhetoric 3*(4), 225–248.

Pennebaker, J. W., R. L. Boyd, K. Jordan, and K. Blackburn (2015). The development and psychometric properties of liwc2015. *UT Faculty/Researcher Works*.

Pestian, J. P., P. Matykiewicz, M. Linn-Gust, B. South, O. Uzuner, J. Wiebe, K. B. Cohen, J. Hurdle, and C. Brew (2012). Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights 5*(Suppl 1), 3.

Phillips, W. and E. Riloff (2002, July). Exploiting strong syntactic heuristics and co-training to learn semantic lexicons. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pp. 125–132. Association for Computational Linguistics.

Popescu, A.-M. and O. Etzioni (2007). Extracting product features and opinions from reviews. In *Natural language processing and text mining*, pp. 9–28. Springer.

Punyakanok, V. and D. Roth (2001). The use of classifiers in sequential inference. In T. K. Leen, T. G. Dietterich, and V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13*, pp. 995–1001. MIT Press.

Purver, M. and S. Battersby (2012). Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 482–491. Association for Computational Linguistics.

Qadir, A., P. Mendes, D. Gruhl, and N. Lewis (2015). Semantic lexicon induction from twitter with pattern relatedness and flexible term length. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Qadir, A. and E. Riloff (2012, 7-8 June). Ensemble-based semantic lexicon induction for semantic tagging. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, Montréal, Canada, pp. 199–208. Association for Computational Linguistics.

Qiu, G., B. Liu, J. Bu, and C. Chen (2011). Opinion word expansion and target extraction through double propagation. *Computational linguistics 37*(1), 9–27.

Rentoumi, V., G. Giannakopoulos, V. Karkaletsis, and A. G. Vouros (2009). Sentiment analysis of figurative language using a word sense disambiguation approach. In *Proceedings of the International Conference RANLP-2009*, pp. 370–375. Association for Computational Linguistics.

Riloff, E. and R. Jones (1999). Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, AAAI '99/IAAI '99, Menlo Park, CA, USA, pp. 474–479. American Association for Artificial Intelligence.

Riloff, E., A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang (2013, October). Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, pp. 704–714. Association for Computational Linguistics.

Riloff, E. and J. Shepherd (1997). A corpus-based approach for building semantic lexicons. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, pp. 117–124.

Riloff, E. and J. Wiebe (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp. 105–112. Association for Computational Linguistics.

Riloff, E., J. Wiebe, and T. Wilson (2003). Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, Stroudsburg, PA, USA, pp. 25–32. Association for Computational Linguistics.

Roark, B. and E. Charniak (1998). Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, ACL '98, Stroudsburg, PA, USA, pp. 1110–1116. Association for Computational Linguistics.

Roberts, K., M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu (2012, May). Empatweet: Annotating and detecting emotions on twitter. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, Istanbul, Turkey, pp. 3806–3813. European Language Resources Association (ELRA).

Rumbell, T., J. Barnden, M. Lee, and A. Wallington (2008). Affect in metaphor: Developments with wordnet. In *Proceedings of the AISB Convention on Communication, Interaction and Social Intelligence*, Volume 1, pp. 21.

Sam, G. and H. Catrinel (2006). On the relation between metaphor and simile: When comparison fails. *Mind & Language 21*(3), 360–378.

Schumaker, R. P., Y. Zhang, C.-N. Huang, and H. Chen (2012). Evaluating sentiment in financial news articles. *Decision Support Systems 53*(3), 458–464.

Sharma, R., M. Gupta, A. Agarwal, and P. Bhattacharyya (2015). Adjective intensity and sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Shutova, E. (2010). Automatic metaphor interpretation as a paraphrasing task. In *Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 1029–1037. Association for Computational Linguistics.

Shutova, E., L. Sun, and A. Korhonen (2010). Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 1002–1010. Association for Computational Linguistics.

Sopory, P. and J. P. Dillard (2002). The persuasive effects of metaphor: A meta-analysis. *Human Communication Research 28*(3), 382–419.

Speer, R. and C. Havasi (2013). Conceptnet 5: A large semantic network for relational knowledge. In *The Peoples Web Meets NLP*, pp. 161–176. Springer.

Staiano, J. and M. Guerini (2014). Depeche mood: a lexicon for emotion analysis from crowd annotated news. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 427–433. Association for Computational Linguistics.

Stone, P., D. C. Dunphy, M. S. Smith, and D. Ogilvie (1968). The general inquirer: A computer approach to content analysis. *Journal of Regional Science 8*(1), 113–116.

Strapparava, C. and R. Mihalcea (2007). SemEval-2007 Task 14: Affective Text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*.

Strapparava, C. and R. Mihalcea (2008). Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pp. 1556–1560. ACM.

Strapparava, C., A. Valitutti, et al. (2004). Wordnet affect: an affective extension of wordnet. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Volume 4, pp. 1083–1086.

Suttles, J. and N. Ide (2013). Distant supervision for emotion classification with discrete binary values. In *Computational Linguistics and Intelligent Text Processing*, pp. 121–136. Springer.

Thelen, M. and E. Riloff (2002). A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, Stroudsburg, PA, USA, pp. 214–221. Association for Computational Linguistics.

Thomas, B., K. Dhanya, and P. Vinod (2014). Synthesized feature space for multiclass emotion classification. In *Proceedings of the 2014 First International Conference on Networks & Soft Computing (ICNSC)*, pp. 188–192. IEEE.

Titov, I. and R. McDonald (2008, June). A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*, Columbus, Ohio, pp. 308–316. Association for Computational Linguistics.

Tokuhisa, R., K. Inui, and Y. Matsumoto (2008). Emotion classification using massive examples extracted from the web. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pp. 881–888. Association for Computational Linguistics.

Tumasjan, A., T. O. Sprenger, P. G. Sandner, and I. M. Welpe (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM 10*, 178–185.

Utsumi, A. and Y. Kuwabara (2005). Interpretive diversity as a source of metaphor-simile distinction. In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, pp. 2230–2235.

Veale, T. (2012). A context-sensitive, multi-faceted model of lexico-conceptual affect. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pp. 75–79. Association for Computational Linguistics.

Veale, T. and Y. Hao (2007). Learning to understand figurative language: from similes to metaphors to irony. In *Proceedings of Cognitive Science*.

Veale, T., E. Shutova, and B. B. Klebanov (2016). Metaphor: A computational perspective. *Synthesis Lectures on Human Language Technologies 9*(1), 1–160.

Vogt, C. C. and G. W. Cottrell (1999). Fusion via a linear combination of scores. *Information retrieval 1*(3), 151–173.

Volkova, S., T. Wilson, and D. Yarowsky (2013, August). Exploring sentiment in social media: Bootstrapping subjectivity clues from multilingual twitter streams. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia, Bulgaria, pp. 505–510. Association for Computational Linguistics.

Vu, T. H., G. Neubig, S. Sakti, T. Toda, and S. Nakamura (2014). Acquiring a dictionary of emotion-provoking events. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pp. 128–132. Association for Computational Linguistics.

Wallington, A., R. Agerri, J. Barnden, M. Lee, and T. Rumbell (2011). Affect transfer by metaphor for an intelligent conversational agent. In *Affective Computing and Sentiment Analysis*, pp. 53–66. Springer.

Wang, W., L. Chen, K. Thirunarayan, and A. P. Sheth (2012). Harnessing twitter" big data" for automatic emotion identification. In *Proceedings of Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pp. 587–592. IEEE.

Wang, X., F. Wei, X. Liu, M. Zhou, and M. Zhang (2011). Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1031–1040. ACM.

Weiner, E. J. (1984). A knowledge representation approach to understanding metaphors. *Computational linguistics 10*(1), 1–14.

Wilson, T., P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan (2005). Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, Stroudsburg, PA, USA, pp. 34–35. Association for Computational Linguistics.

Wilson, T., J. Wiebe, and P. Hoffmann (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pp. 347–354. Association for Computational Linguistics.

Wu, W., H. Li, H. Wang, and K. Q. Zhu (2012). Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pp. 481–492. ACM.

Xu, Y., Y. Wang, J. Liu, Z. Tu, J.-T. Sun, J. Tsujii, and E. Chang (2012). Suicide note sentiment classification: a supervised approach augmented by web data. *Biomedical informatics insights 5*(Suppl 1), 31.

Yang, C., K. H.-Y. Lin, and H.-H. Chen (2007a). Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pp. 133–136. Association for Computational Linguistics.

Yang, C., K. H.-Y. Lin, and H.-H. Chen (2007b). Emotion classification using web blog corpora. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, WI '07, Washington, DC, USA, pp. 275–278. IEEE Computer Society.

Zhu, X., S. Kiritchenko, and S. Mohammad (2014, August). Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, pp. 443–447. Association for Computational Linguistics and Dublin City University.