Emotion Combination in Social Media Comments as Features for Sarcasm Detection

Fernando H. Calderon
Social Networks and Human Centered Computing Program, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica
128 Academia Road Sec.2 Nangang
Taipei, Taiwan 115
fhcalderon87@gmail.com

Po Chen Kuo
Department of Computer Science, National Tsing Hua University
GuangFu Rd. Sec.2 No.101
Hsinchu, Taiwan 30013
chuck82521@gmail.com

Huang Yen-Hao
Institute of Information Systems and Applications, National Tsing Hua University
GuangFu Rd. Sec.2 No.101
Hsinchu, Taiwan 30013
yenhao0218@gmail.com

Yi-Shin Chen*
Institute of Information Systems and Applications, National Tsing Hua University
GuangFu Rd. Sec.2 No.101
Hsinchu, Taiwan 30013
yishin@gmail.com

ABSTRACT

Emotions or mood play a significant part in social media users' interactions. This, in turn, has filled these platforms with opinionated content. Due to the nature of the data, the variety of platforms, and dynamic online user behavior, approaches to make better use of this data face many issues. It remains a challenge to properly obtain a reliable emotional status from a user when posting a comment. Sarcasm is one of the behavioral factors that adds difficulty to this task. Sarcastic expressions have an impact on both emotional expression and perception. The lack of explicit labels presents an additional challenge. This work introduces a methodology that explores the overlap of emotions in a text as an inherent characteristic of sarcastic expressions. To obtain such emotion labels, we leveraged a semi-supervised emotion-detection system trained using Facebook reactions and comments. Textual information was converted into a graph from which linguistic patterns were extracted and used as features. This pattern extraction method enabled multi-lingual usage and represented context in expressions more efficiently. More than 1 million English and Chinese comments from over 62,000 public Facebook page posts have been collected and processed; experiments conducted show acceptable performance metrics.

KEYWORDS

sentiment analysis, emotion detection, social media, sarcasm detection

*Associate Professor

WISDOM '19, August 04, 2019, Anchorage, Alaska,

1 INTRODUCTION

Social media platforms have long been regarded as a rich data source, especially since it is possible to understand opinions and emotions expressed in them toward a particular subject or object. Facebook has been for some time now one of the leading online social networks [29]. Within the platform we find Facebook "pages" which are in essence official accounts for an individual, media, or organization. Pages tend to receive more comments and reactions than regular user accounts. Reactions allow users to express a series of emotions, in addition to the "like" button. This has made Facebook page posts a place laced with opinions, emotions, and sarcasm. It is important to understand how these interact together. It can be said that sarcasm is used to invert emotions, but conversely, can the interaction of certain emotions be an indicator of sarcasm?

When trying to detect sarcasm, context plays an important role, which implies an understanding of several factors in the setting of a comment [2, 11]. The topic, background of the user, background of the receiver, and emotions conveyed can provide some insight when determining if a comment is sarcastic or not. The problem is that not all of these factors are available when attempting to train an algorithm to detect sarcasm. State-of-the-art text classification methods face additional challenges, such as lack of annotated data. Due to the nature of sarcastic posts they are not usually explicitly labelled. Other methods that use manually selected features are hard to maintain. Moreover, even when these contextual cues are presented to a human reader, sarcasm may still be hard to detect. It is thus helpful to look into other features that can provide clues or hints on the presence of sarcasm.

Some of these features may be related to the ways of interacting on social media platforms themselves. For example, one is more likely to find a sarcastic reply or comment than a sarcastic post; that is, a response to try to outsmart an original post [18]. Understanding this dynamic can help us better identify data resources where sarcasm is more prone to appear. On the other hand, features that are implicit to the textual content gain importance since the data to be collected must at least contain a comment. Emotions are strongly tied to sarcastic expressions since they may be contained in them or affect the perception of the reader.

Figure 1: Example of a sarcastic comment on a news post.

The example interaction in Figure 1 highlights the mentioned phenomena. It can be observed that the sarcastic comment appears in the form of a reply. Within its content there are three statements with different polarities. The closing statement, "Have a great day!" has a positive connotation; however, the other components appear more negative. This might pose a challenge for emotion-recognition systems. Nevertheless, we believe this variation of emotions can be used as a feature for sarcasm classification.

This work tries to make use of the inherent characteristics of the Facebook platform with the objective of developing a system for emotion detection based only on the content extracted without the need for external knowledge. In essence, our system will leverage crowdsourcing to achieve this goal. First, it uses the intersection of reaction clicks and comments as a fuzzy labeling technique where the reactions become the labels of the comments to train our emotion classifier. The emotion classifier is based on an unsupervised graph-based approach that extracts linguistic patterns to be used as features. Such features represent word sequences that highlight context and intention expressed in text. It is then explored if multiple emotions identified within a comment can be an indication of sarcasm. Experiments have been performed for both English and Chinese comments and an extended evaluation for English is presented.

To the best of our knowledge, the self-reported reactions have not been previously used as emotion signals for labeling, nor has the possibility of detecting sarcasm from this kind of classifier been explored before. The work presents the following contributions:

- A semi-supervised emotion classifier using Facebook comments and reactions(as labels) for training.
- An empirical analysis to discover which emotion combinations can hint sarcasm.
- An extensive experimental evaluation on the task of sarcasm detection highlighting the performance of the proposed emotion based method compared to other baselines in text classification.
- The described methods can be implemented for multiple languages as the experiments show.

2 RELATED WORK

2.1 Sentiment Analysis on Social Media

Online social media platforms have increasingly attracted more interest for the sentiment and emotions expressed in their users' opinions. This has led to the inclusion of several explicit means to reflect such sentiments in a comprehensive, user-friendly, and collectible way. Several works have focused on using these signals as noisy labels for sentiment classification [1, 5, 7, 9, 28, 30]. The work by Go et al. [7] evaluated the performance of popular machine learning algorithms when using emoticons in tweets as labels for training via distant supervision. In a similar way, Davidov et al. [5] leveraged Twitter features for sentiment learning and not only considered emoticons as labels but also added hashtags.

Previous works confirmed at the time that social media could provide not only data but annotated data that could avoid the time- and resource-intensive task of manual annotation. This advantage was further explored by Zhao et al. [30] using data from a different platform (Weibo) and language (Chinese). Their system, MoodLens, mapped 95 emoticons into four sentiment classes and became one of the pioneering tools for sentiment analysis from short texts in Chinese. Another study using the Weibo platform performed sentiment correlation to determine which of two emotions–anger and joy–is more influential in a social network [6]. Lipsman et al. [13] focused uniquely on the number of "likes" in a post to determine what kind of repercussions this click behavior had from the perspective of brands and fans.

Following a similar trend, the work by Hu et al. [9] studied the use of emotion signals not only as labels for training but also as an active part in unsupervised learning models for sentiment analysis. Hashtags on their own have also been used for similar tasks. Argueta et al. [1] used hashtags for distant supervision on unsupervised methods to collect writing patterns that can be correlated to emotions. It has been found that using emoticons or hashtags as labels can lead to some errors. This served as motivation for Wang et al. [28], who proposed a method for "de-noising" the obtained labels.

Despite the availability of multiple online social networks, most of the related work has been focused on Twitter. Ortigosa et al. [20] were among the first to perform sentiment analysis on Facebook. Their application, SentBuk, tries to help e-learning systems by providing sentiment information of users through their posts. The achieved performance shows that Facebook data can also be used for sentiment-related tasks.

Recent years have witnessed the development of algorithms that deliver a very high performance on sentiment related tasks. VADER, the rule-based model developed by Hutto and Gilbert [10], aims to make the most out of sentiment lexicons combined with machine learning algorithms. Deep convolutional neural networks have also participated significantly in sentiment classification tasks. The work by Poria [22] presents a model for multi-modal classification of short sentences based on features extracted from text. As highlighted by Liu [14] and Cambria [4], however, there are several factors affecting sentiment-related topics, many of which have not been thoroughly explored. For instance, Volkova et al. [27] explored the impact of demographic language variations when attempting multilingual sentiment analysis and made clear how this can be an issue. The context in which opinions are expressed is also of high importance, as studied by Muhammad et al. [19]. Their work explains that significant variations in modeling are required depending on the social media genre being studied. The impact of innate human responses, such as sarcasm, is also one of the factors requiring in-depth exploration.

2.2 Sarcasm Detection

Sarcastic expressions are a natural product of humor improvisation and have naturally proliferated in online social media. With them, they bear a lot of trouble for mining tasks due to the uncertainty and ambiguity they bring to expressions. If it is hard for humans to define and identify sarcasm, it is even harder to teach a computer to do so. Nevertheless, the research community has attempted to achieve this.

Maynard and Greenwood [17] highlight the importance of understanding the impact of sarcasm in sentiment analysis. Reyes et al. [25] first attempted to identify humor and irony on Twitter posts, as this could provide some insights to sarcastic expressions. Based on textual features and leveraging on the hashtags *#humor* and *#irony*, they developed a system to identify "figurative language". Bamman and Smith [2] believe that sarcasm is a highly contextual phenomenon and that extra-linguistic information is required for its detection. They considered lexical cues and their corresponding sentiment as contextual features in their study.

Rajadesingan et al. [24] go beyond these affirmations and claim that behavioral traits are also intrinsic to users expressing sarcasm. They developed a model for sarcasm detection based on the analysis of past tweets paired with behavioral and psychological studies. Riloff et al. [26] attempted to identify instances in which a situation and its subsequent reaction have opposing sentiment polarities. They used this as a clue to identify sarcastic expressions using bootstrap learning methods. Gonzalez-Ibanez et al. [8] also experimented with the sentiment polarity in Twitter messages and the presence of sarcasm transforming this polarity. Their work used lexical and pragmatic features to train a machine learning system to identify these utterances. Lexical features were also used by Bharti et al. [3] in developing their parsing-based lexicon-generation algorithm to detect sarcasm on Twitter.

Sarcasm detection has also been attempted in other languages. The work by Lunando and Purwarianti [16] first performed sentiment classification on short texts from Indonesian social media. It then considered two other features: negativity information and the number of interjection words to perform sarcasm detection through machine learning algorithms. Liebrecht et al. [12] built a Twitter-based corpus by collecting tweets containing the hashtag *#sarcasm* and trained a machine learning classifier. The data used was in Dutch, but still showed that sarcasm is often signaled by intensifiers and exclamation marks. Tweets in English and Czech were studied by Ptacek et al. [23] to develop a language-independent approach borrowing features across languages. The work by Liu [15] explores sarcasm detection in Chinese text, primarily focusing on the issue of imbalanced data and proper feature selection, which was evaluated through multiple classifiers. The dataset used contained simplified Chinese characters; to the best of our knowledge our work is the first to address this problem for traditional Chinese characters.

3 METHODOLOGY

3.1 Overview

Borrowing from some of the approaches mentioned in the related work, emotion classification was first performed on short texts. The training data and labels were obtained from the intersection between reaction clicks and comments from users corresponding to those reactions. The classifier returned the two most likely labels corresponding to a text. These two labels then underwent feature evaluations that helped determine if a comment was sarcastic or not. A flowchart of the method is presented in Figure 2.

3.2 Data Collection

One of the key features of this work is the exploitation of embedded characteristics of the Facebook platform, the first being their "pages" feature. Facebook pages are official accounts of varied types of sources, popular personalities, organizations, and media outlets. Our methodology takes particular advantage of the official pages of news media. The implemented emotion classification algorithm requires objective and subjective data in its development. By using pages from news media, objective texts can be obtained from news posts while subjective texts can be obtained from user comments. It is assumed that comments on these articles are usually highly opinionated, sometimes biased, and predominantly subjective.

The second key element to be used is Facebook "reactions." Since the beginning of 2016, the traditional "like" button was replaced by a more variety-aware option called "reactions." Reactions are emoji-based expressions that allow a user to express their sentiment toward a post. The set of available reactions include "Love", "Haha", "Wow", "Sad" and "Angry"[1]. It was identified that many of the users who react to posts also have a tendency to comment on them. The proposed data collection approach was to find the intersection between reaction clicks and comments that would enable a match between a user comment and its corresponding reaction. This not only allowed a filtering process to build a collection but also guaranteed that there would be a self-reported emotion assigned to every comment, resulting in an automatic annotation.

The previously mentioned characteristics allowed for a collection of objective news data and subjective comments data, the latter of which was paired with emotion labels. This met the requirements for the implementation of the pattern-based emotion classifier to be used in this work.

3.3 Reaction-Based Emotion Classification

To obtain the emotion labels to be used as features for sarcasm detection, a multi-lingual pattern-based emotion classifier was implemented [1]. The approach first constructed a graph densely populated with subjective expressions from short texts. It then extracted linguistic-based patterns of expression, which could be weighed across the collected range of emotions. The content-driven,

[1]Even though some of these reaction emojis are not emotions in their strict sense (as listed in Plutchik's wheel of emotions [21]), they may nevertheless provide some insight into a user's sentiments.

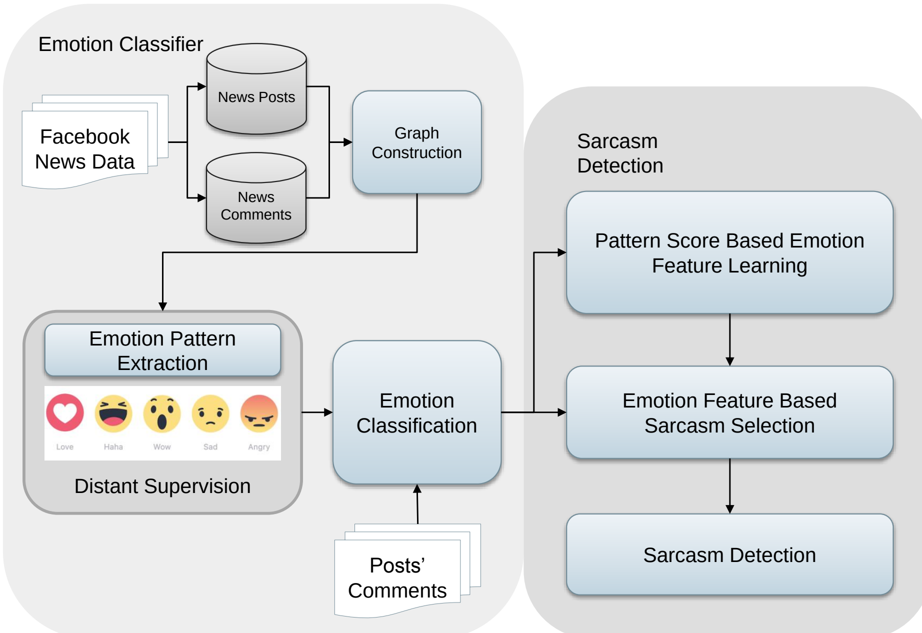

Figure 2: Methodology flowchart.

pattern-based methodology allowed for replication across different languages.

3.3.1 Graph Construction. The data obtained, as defined in the previous section, was converted into graph form for further manipulation. The nodes in the graph correspond to words, and the edges denote the co-occurrence between the connected words. The order in which words appear was also considered in the co-occurrence and hence reflected in the direction of the edges. The graph obtained from the news posts intuitively contains more factual, objective expressions, while the graph from the comments is more subjective and opinionated.

Figure 3 shows examples of graphs constructed from comments (i.e., subjective) in Chinese and English.

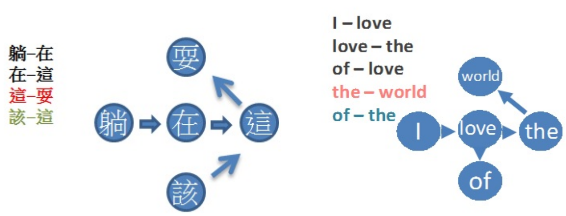

Figure 3: Examples of subjective graphs from Chinese and English comments.

The next step was to obtain a set of expressions that were highly subjective so that they would have a stronger link to emotions. With this purpose, a reduction was performed on the comments graph by removing terms that were highly dominant in the news graph. This procedure reduces the objective components present in the comments graph, resulting in a highly subjective graph. An example of graph reduction is shown in Figure 4 for both Chinese and English graphs.

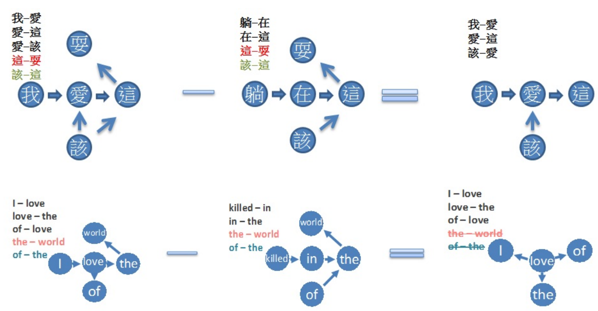

Figure 4: Examples of graph reduction for Chinese and English graphs.

The graph construction process for Chinese text posed an additional difficulty since it required more steps in its pre-processing, particularly in word segmentation. In Chinese, a word can be composed generally by one, two, or three characters. Characters may have one meaning when they appear alone, and they can mean something totally different when paired with another character. It is therefore highly important to perform appropriate character segmentation and sense disambiguation before proceeding to build the graph. In order to segment Chinese characters, the following hierarchical combination was performed:

(1) Combine the characters into two-, three-, and four-character words and calculate their frequency in the dataset.
(2) Perform an initial reduction on three-character words by subtracting the frequency of four-character words that contain them.

(3) Perform an additional reduction on two-character words by subtracting the frequency of three- and four-character words that contain them.

After the previous procedure was complete, frequent words were filtered into two-, three-, and four-character words using an arbitrary frequency threshold.

3.3.2 Emotion Patterns. Repetitive instances of sequences in the graph with $length = 2$ or $length = 3$ sharing one or two words became patterns. Since the graph is filled with subjective expressions, the assumption was that the obtained patterns are expressions that denote a high level of emotion. It is also important to determine which emotion a pattern is more likely to be expressing.

Definition 3.1. An element e is a word or a sequence of symbols (,.?!, etc).

Definition 3.2. A pattern P_i is a sequence of two or three elements.

$$P_i = [e_1, e_2, e_3] \forall P_i \in \mathcal{P} \quad (1)$$

The obtained emotion patterns were then paired to our set of labels through distant supervision. Once the set of patterns was obtained and a set of comments with their corresponding emotion labels collected, we counted how many instances of the patterns were in the corpus. Through probabilistic analysis, it was determined in which particular emotion label a certain pattern was more predominant.

Definition 3.3. An Emotion Degree $ED(emo, p)$ is a score representing how a pattern is related to a specific emotion.

$$ED(emo, p) \longmapsto ed, ed \in \mathbb{R}^+ \quad (2)$$

Definition 3.4. An emotion emo is defined in a set of 5 emotions. All the patterns have 5 Emotion Scores(ES) each with a corresponding emotion.

$$emo \in Emotion\{Haha, Angry, Sad, Love, Wow\}$$

As a result, every emotion will contain the same patterns, but ranked in a different order and weighed by Emotion Degree ED that depends on their frequency, uniqueness, and diversity.

Definition 3.5. Pattern Frequency (PF)

Pattern Frequency $PF(emo, p)$ represent the frequency of an emotion pattern p in a collection of social data related to emotion emo.

$$PF(emo, p) = log(f(p, emo) + 1) \quad (3)$$

Definition 3.6. Inverse Emotion Frequency (IEF)

The Inverse Emotion Frequency $IEF(emo, p)$ is a measurement of how rare or unique a pattern p is across all emotion classes.

$$IEF(emo, p) = \frac{|Emotion|}{|\{emo \in Emotion : f(p, emo) > 0\}|} \quad (4)$$

Definition 3.7. Diversity (DIV)

Diversity $DIV(p)$ considers the number of unique psychology words (denoted as uew that fit the pattern p across all emotion classes emo.

$$DIV(p) = log(uew(p, Emotion)) \quad (5)$$

Table 1: Examples of highly ranked patterns for some emotion labels in English.

Angry	Haha	Wow	Sad
* all haters	* . lol	* . awesome	* so sad
trump is *	happy bday *	a * what	my heart *
what a *	* ! yeah	* user omg	* god bless
* this country	ever !! *	!!!! * !	prayers for *
people are *	looks so *	* !!! how	. rip *

Finally, the emotion degree that shows how important a pattern is in an emotion class is obtained by the equation below.

$$ED(emo, p) = PF(emo, p) \times IEF(emo, p) \times DIV(p) \quad (6)$$

Table 1 contains examples of extracted patterns that are ranked high and thus very representative of the corresponding labels. It is worth noting that the corpus was generated from a crawl of Facebook pages of news media, so at the time of the crawl, these were rich in political content, international conflicts, etc. This can be particularly evident for Angry and Sad, where the topic related to the corresponding comments that generated these patterns can be deduced. Other categories also have particular characteristics. For example, in the Wow emotion, there is a presence of question words such as "what" and "how" which can be indicators of surprise.

The presence of the wildcard (*) in the patterns is also worth noticing. The wildcard takes the place of a word that can elicit a high degree of sentiment. These words are replaced by this token so that any word that is used in the same way can be matched by these patterns. For instance, the pattern "people are *" could match "people are dumb" or "people are stupid," both denoting an angry expression; the usage of the wildcard thus allows matching both examples to the same pattern. A pattern's ability to capture many different instances is what was referred to previously as diversity. Altogether these patterns can be interpreted as an extension of word embeddings, with the inclusion of context from surrounding words, adaptability to different domains, and resilience to unseen terms through the integrated wildcard.

3.3.3 Emotion Classification. The classification process then takes a new unlabeled comment and, through a matrix multiplication procedure, evaluates it with the patterns and ranks within the labels. The process first determines which patterns are present in the post. It then proceeds to calculate the score of how likely a text is to belong to a class, depending on the score and ranking of the patterns it contains. As a result, the system returns a scored and ranked list of the emotion labels based on the likelihood of the new comment belonging to them. For practical purposes, the top two results are considered as the labels for the input text. These two labels are then evaluated to see if they can provide insights into the presence of sarcasm.

3.4 Sarcasm Detection

Sarcasm is a highly context-dependent reaction–it is usually not planned for, but initially depends on previous information. The post-then-comment scenario from which the data was crawled provides a kind of interaction that may favor this behavior. For example, a user first reads a news post and, depending on his or her opinions toward the topic, he or she might decide to first react to it in a peculiar way and then provide a sarcastic reply to the post.

Another characteristic of sarcasm–and one that has troubled the sentiment analysis community–is the reverting of an emotion from the perceptive point of view. This is the typical use of positive statements when actually having a negative point of view. If the receiver is not aware of the state of humor or behavioral traits of the sender, then the message may be perceived as positive, while the intention may have been negative. This poses a significant challenge to automatic emotion classification systems, since they cannot be aware of these particular behavioral traits. Our methodology tries to make use of this combination of emotions as a feature for sarcasm detection, as explained in the following section.

3.4.1 Sarcasm Candidate Filtering. Based on the aforementioned user behavior, this stage initially determines if a comment is at all eligible for containing sarcasm. After performing emotion classification for a large set of comments, a particular case arose in which many of the results consisted of opposing sentiment labels, specifically Angry and Haha. This is perhaps due to the nature of the data and the presence of internet trolls on these kind of sites, which can lead to a user reacting with laughter to a piece of news that would otherwise generate anger in the majority of the population. Nevertheless, this also relates to sarcastic behavior.

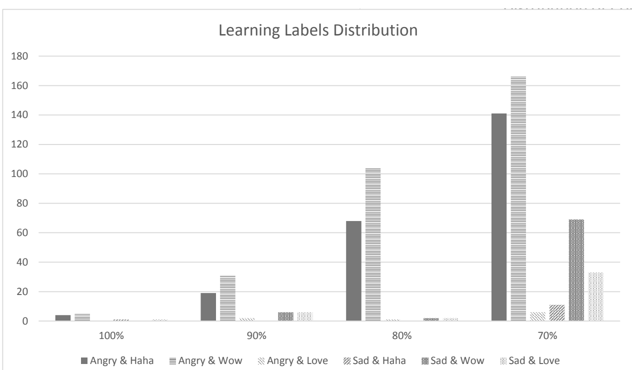

Figure 5: Example of emotion combination Learning Result for English comments.

Considering this behavior, every short document with opposite emotions was considered a candidate for sarcasm detection. To find the combination of emotions that can indicate sarcasm, a machine learning method was used to quantify this possibility. The proposed method made use of combined emotions as a feature for sarcasm detection, as explained in the following process.

- Convolutional Neural Network
  - Input Matrix

    Since every emotion will contain the same patterns weighted by a score and ranked in different order accordingly, the Pattern Scores *PS* of the comment in each emotion are considered as the input matrix of the Convolutional Neural Network.

    Every comment evaluated will generate an input matrix as follows:

$$\begin{array}{cc} & \begin{matrix} PatternScore_1 & \dots & PatternScore_n \end{matrix} \\ \begin{matrix} Emotion1 \\ Emotion2 \\ Emotion3 \\ Emotion4 \\ Emotion5 \end{matrix} & \begin{pmatrix} 300 & \dots & 2000 \\ & & \\ & & \\ & & \\ & & \end{pmatrix} \end{array}$$

    The Pattern Score is the ranking of the pattern in each emotion multiplied by the frequency of the pattern in the comment .Here we consider n patterns in every emotion where n is experimentally defined.
  - Training Prediction

    When training the model, the training prediction value is used as the observation source. It is then calculated in how many of the iterations is a sarcastic comment correctly identified, in parallel this highlights those that can be correctly learned by our model. This value is termed the correct training rate. Since sarcasm has the characteristic of flip of emotion, the combinations of opposite emotions that allow a correct prediction are considered. By calculating the emotion combination results for the range 100% to 70% correct training rate, we can extract the specific combinations which represent more precise indicators for sarcastic comments. For example, by observing the correct learning rate at 70% we can identify which 2 combinations of emotions where more useful for the classification as illustrated by Figure 5.

After performing this evaluation, it was observed that the emotion pairs Angry & Haha and Angry & Wow are the most useful as learning features for sarcasm detection. It was therefore determined that every short document with these two emotion labels resulting from the classifier was thus considered a candidate for sarcasm detection.

3.4.2 Sarcasm Labeling. It was observed that just the presence of the two specific emotion labels was not directly an indicator of sarcasm. There is a dependency on the distance between these two initial labels. If the top label is very dominant compared to subsequent labels, there is less chance that it is a sarcastic instance. On the other hand, if the two top labels have similar scores and opposing sentiment, there is a higher probability that it is a sarcastic comment. Once a sarcasm candidate was received, two measurements between its two emotion labels were obtained. These values were the determining factor in deciding if a comment was sarcastic or not, and are specified by Definition 3.8 and Definition 3.9.

Definition 3.8. Distance Ratio Measurement

To make sure there is not only one specific emotion, the difference of emotion score of emotion 1 and 2 with emotion 2 and 3 is measured. The difference of emotion score of emotion 1 and 2 by emotion 2 and 3 is then divided. The resulting value of the measurement needs to be greater or equal than x_1, and less than or equal than x_2 where x_1 and x_2 are experimentally defined.

$$x_2 \geq \frac{Score(emotion_3) - Score(emotion_2)}{Score(emotion_2) - Score(emotion_1)} \geq x_1 \quad (7)$$

Definition 3.9. Score Ratio Measurement

To make sure there is not only one specific emotion, the emotion score ratio of emotion 1 and 2 with emotion 2 and 3 is calculated. The value of emotion score ratio of emotion 1 and 2 needs to be greater or equal than y_1, and the value of emotion score ratio of emotion 2 and 3 needs to be greater or equal than y_2 where y_1 and y_2 are experimentally defined.

$$\frac{Score(emotion_3)}{Score(emotion_2)} \geq y_1, \frac{Score(emotion_2)}{Score(emotion_1)} \geq y_2 \quad (8)$$

The first guarantees that there is not only one emotion by evaluating distance between scores. The second one is for normalization, since the ratio from one emotion to the next can be a better indicator than just the distance. It is worth mentioning that the thresholds vary across languages, perhaps due to cultural differences and language expression. It was found, for example, that Chinese posts tend to contain one dominant emotion, with an outstanding score. Therefore the range must adjust to this characteristic.

4 EXPERIMENTS AND RESULTS

4.1 Data

The collected posts came from a variety of public Facebook pages belonging to news media outlets in both Chinese and English; each dataset was evaluated separately. A total of 62,248 posts were crawled, together with the comments and reactions they contained. Around 46,253 posts with approximately 3 million comments corresponded to Chinese data, crawled during the period between June 1 and July 31, 2016. The remaining 15,995 posts with approximately 7 million comments were in English and were obtained during the period between October 1 and November 30, 2016. After the comments were collected, they were matched with a corresponding reaction chosen by the user. Table 2 presents the total counts of comments overlapped to a particular reaction for both English and Chinese datasets. These sets of comments, with their self-reported annotations, were used to train the system.[2]

Table 2: Counts of collected comments per corresponding emotion for both English and Chinese.

Overlapped Emotion	Chinese Comments	English Comments	Total
Angry	167,692	206,994	374,686
Haha	79,444	162,149	214,593
Wow	38,433	61,720	100,153
Sad	28,271	102,264	130,535
Love	34,019	300,600	334,619
Total	347,859	833,727	1,181,586

The comments in Chinese were in traditional Mandarin characters from predominantly Taiwanese news media. It is interesting to notice how the English posts had a much higher comment density. The percentage that every reaction represents in the datasets is presented in Figure 6. It can be observed that the distribution for both Chinese comments and English comments have similar behaviors for the reactions in the middle, but very opposing distributions for Angry and Love. This kind of distribution can provide some insights into how different groups interact with the platform. Furthermore, this can lead to a deeper study on the differences or similarities in interactions based on cultural or language backgrounds. However, it must be made clear that this behavior may not be universal and is probably dependent on the time period crawled and the trending news of that period.

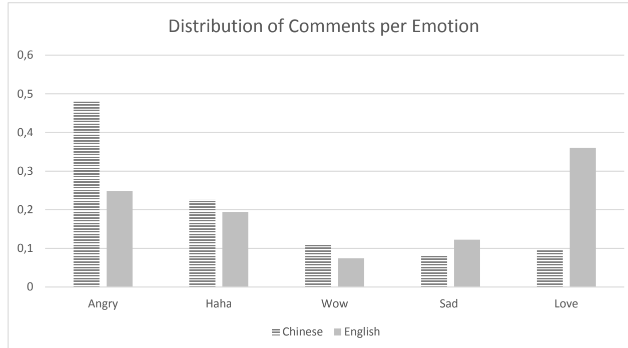

Figure 6: Distribution of comments across emotions for English and Chinese.

Another factor to take into account is that the data comes from news media posts; therefore, these posts may have a tendency to elicit certain reactions more than others. This can explain why Angry has a significant share in both languages, while Wow and Sad are not as common, perhaps due to the nature of the news shared.

Separate sets of data were crawled at different times to perform the evaluation. For evaluation, human annotation was required. The following subsection describes the process for ground truth generation.

4.2 Ground Truth

As previously mentioned, one of the main challenges in the sarcasm-detection task is the difficulty, even for humans, of identifying sarcastic expressions. Annotated testing sets of good quality and consistency were required in order to perform a proper evaluation. Several testing sets were constructed for this purpose: four for the English evaluation and three for the Chinese. The inter-annotator consistency of the datasets is measured by Fleiss' kappa. A good inter-annotator score guarantees the quality of the testing sets, while maintaining this quality across sets is an indicator of consistency. The details of the testing sets are provided in Table 3.

The annotation task required the annotators to label a comment as being sarcastic or not. The comments for both the English and Chinese sets were collected from Facebook. The texts were collected randomly from several time periods to avoid any particular bias to a specific news event or season.

[2] The collected sets and corresponding labels can be made available upon request.

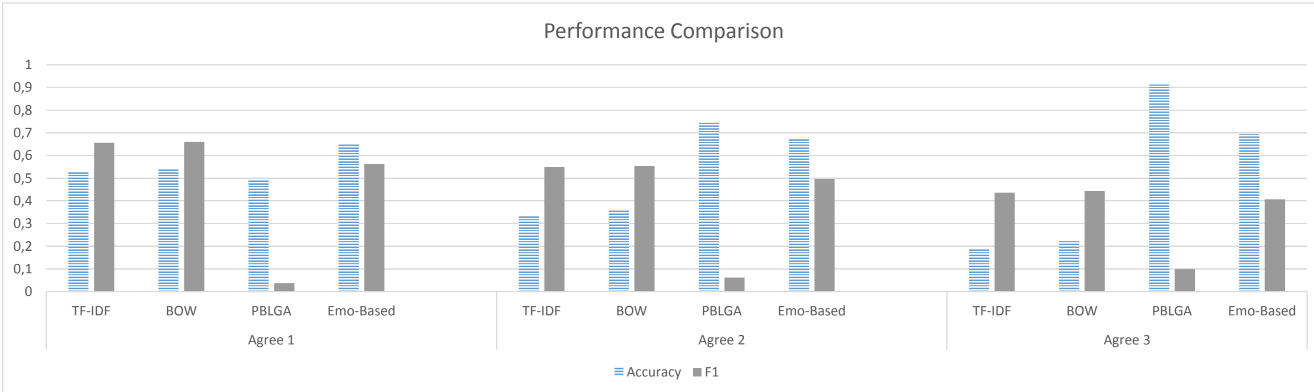

Figure 7: Sarcasm classifier performance comparison by Accuracy and F-Measure for English experiments on the Amazon Turk Test set.

The annotators for the English and Chinese Test 1, 2, and 3 sets were university students between 22 and 29 years old and native speakers of the language being evaluated. The subjects were familiar with the social media platforms and sarcastic posting behavior. Not all annotators evaluated all of the sets, as shown in the "#of Annotators" column in Table 3. Different combinations of annotators worked on different sets, but, as observed, they maintained good Fleiss' kappa scores. According to the suggested interpretation, all sets achieved at least substantial agreement (0.61-0.80).

The Test Turk set came from a task submitted to Amazon Mechanical Turk, where every text was rated by three annotators. Additionally all annotators were asked to provide a degree of intensity, which is not used in this work but might be useful in the future. The task contained a few manually inserted comments regarded as definitely sarcastic and definitely not sarcastic to verify if the annotators could perform the evaluation correctly.

4.3 Evaluation

4.3.1 English Data Test. To evaluate the performance of our Sarcasm Classifier for English texts, three comparison methods were implemented. The first two were text classification baselines trained with a corpus related to the topic at hand using features based on Term Frequency-Inverse Document Frequency(TF-IDF) and Bag of Words (BOW), respectively, to train Naïve Bayes classifiers. Both classifiers were trained using 2400 short documents, 1200 being sarcastic and the other 1200 with no presence of sarcasm. The third comparison method was an implementation of the Parsing-Based Lexicon Generation Algorithm (PBLGA) method developed by Bharti et al. [3]. This method was trained with 40,000 short documents containing the hashtag #sarcasm as indicated by the referenced work. The method introduced in this work will be referred to as Emo-Based in the results to be presented.

Table 3: Details of the testing sets to be used for evaluation.

Language	Set	# of Texts	# of Annotators	Fleiss' Kappa
English	Test 1	430	3	0.7426
	Test 2	260	4	0.7391
	Test 3	400	4	0.7563
	Test Turk	720	3	0.7148
Chinese	Test 1	300	5	0.8560
	Test 2	294	6	0.9104
	Test 3	346	6	0.7630

All four English test sets were processed by the four classifiers mentioned in the previous paragraph. Three different levels of agreement were considered to determine the correctness of a classification: **Agree 1** means that the output label of the classifier matched the label of at least one annotator. **Agree 2** required the classifier to match the label of at least two human annotators. Subsequently, **Agree 3** means at least three annotators agreed to a label and so did the classifier. Presumably Agree 1 will contain more texts regarded as sarcasm by the annotation since it only requires one annotator to label it as sarcasm, while Agree 3 will contain fewer cases of sarcasm since three annotators needed to agree on it, giving it a stricter policy.

Figure 7 presents accuracy and F1-score performance for all classifiers across the three described levels of agreement for the

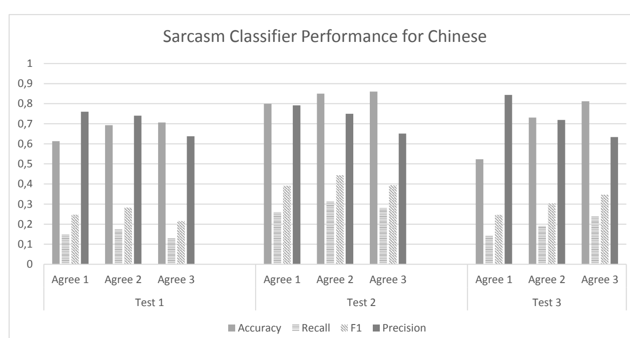

Figure 8: Performance of sarcasm classifier for Chinese data illustrated by Accuracy, Recall, F1-Score and Precision metrics.

Table 4: Performance comparison against other methods for different data sets and varying levels of annotator agreement.

		Agree 1				Agree 2				Agree 3			
Set	Method	*Accuracy*	*F1*	*Recall*	*Precision*	*Accuracy*	*F1*	*Recall*	*Precision*	*Accuracy*	*F1*	*Recall*	*Precision*
Test 1	***Emo-Based***	***0,6023***	0,4896	0,3814	**0,6833**	*0,6535*	0,4494	0,3840	**0,5417**	*0,7047*	0,4227	0,4286	**0,4170**
	PBLGA	0,4907	0,0179	0,0093	0,2500	0,7000	0,0301	0,0160	0,2500	0,8674	0,0000	0,0000	0,1650
	TFIDF	0,5140	0,6579	0,9349	0,5076	0,3186	0,5539	0,9120	0,3977	0,1651	0,4465	0,8776	0,2994
	BOW	0,5209	0,6544	0,9070	0,5118	0,3349	0,5502	0,8800	0,4003	0,1953	0,4463	0,8571	0,3017
Test 2	***Emo-Based***	***0,6000***	0,5000	0,3939	**0,6842**	*0,6769*	0,4891	0,4429	**0,5461**	*0,6615*	0,3528	0,3125	**0,4051**
	PBLGA	0,4923	0,0149	0,0076	0,5000	0,7231	0,0000	0,0000	0,2500	0,8692	0,0000	0,0000	0,1650
	TFIDF	0,5385	0,6512	0,8485	0,5283	0,3692	0,5456	0,8429	0,4033	0,2846	0,4650	0,9063	0,3127
	BOW	0,5423	0,6469	0,8258	0,5317	0,3808	0,5408	0,8143	0,4049	0,3115	0,4679	0,9063	0,3153
Test 3	***Emo-Based***	***0,6075***	0,5341	0,4545	0,6475	*0,6150*	0,4710	0,4370	**0,5108**	*0,6050*	0,3726	0,3538	0,3934
	PBLGA	0,5125	0,0580	0,0303	0,6667	0,6950	0,0480	0,0252	0,5000	0,8300	0,0836	0,0462	0,4433
	TFIDF	0,5075	0,6359	0,8687	0,5015	0,3500	0,5419	0,8487	0,3980	0,2700	0,4712	0,8923	0,3201
	BOW	0,5125	0,6314	0,8434	0,5045	0,3650	0,5387	0,8235	0,4003	0,3000	0,4751	0,8923	0,3238
Turk	***Emo-Based***	***0,6528***	0,5614	0,4444	**0,7619**	*0,6778*	0,4956	0,4389	**0,5690**	*0,6944*	0,4067	0,4038	**0,4096**
	PBLGA	0,4972	0,0372	0,0194	0,4375	0,7444	0,0616	0,0333	0,4063	0,9139	0,0983	0,0577	0,3319
	TFIDF	0,5292	0,6572	0,9028	0,5167	0,3375	0,5485	0,9222	0,3903	0,1903	0,4366	0,9423	0,2841
	BOW	0,5444	0,6605	0,8861	0,5264	0,3611	0,5527	0,9056	0,3977	0,2222	0,4435	0,9423	0,2900

Amazon Turk test set. It can be observed that the Emo-Based method proposed in this work has a more stable performance across levels of agreement, especially regarding accuracy. The TF-IDF and BOW methods perform well in the Agree 1 evaluation since they are content-based methods, and the first level of agreement contains more texts labeled as sarcasm. However, their performance is affected as the ground truth becomes stricter. After reviewing the classification by these methods, we found they are very generous in assigning a sarcasm label since it is determined by the presence of specific terms but does not consider context.

The PBLGA method, on the other hand, did not perform well in terms of F1 score but improved in accuracy with the level of agreement, which was surprising. After looking at the classification output of PBLGA, the opposite situation was found from the content-based methods: PBLGA is very selective on where to label sarcasm, and therefore the fewer cases of sarcasm present in the ground truth the better the accuracy, since it will classify most of the non-sarcasm correctly. Nevertheless it suffers significantly in recall. The Emo-Based sarcasm classifier receives emotions from a pattern-based approach, which can provide more context. Additionally, the candidate filtering process, as well as the distance ratio and score ratio measurements introduced in the methodology, made the classifier less generous–yet at the same time not overly selective–when labeling a text as sarcasm. Complete details of the performance results across all data sets and including additional indicators of recall and precision is presented in Table 4.

The results reflect a more consistent performance from the Emo-Based method across datasets and agreement levels. It can be observed that it had the highest level of accuracy for Agree 1 on all data sets; more importantly, this performance is maintained despite the ground truth becoming stricter. The proposed method also had the better precision across the board, reflecting a good performance of the filtering process defined in the methodology. Other methods tend to lean to one class, which in the case of TF-IDF and BOW favors their recall. It can also be observed that the PBLGA method presented no score for F1 or recall in some instances. When analyzed further it was found that PBLGA labeled very few texts as sarcasm, and as the agree level increased, fewer instances of sarcasm remained. The 0 scores in the table are a result of there being no matching between these scarce labels from the classifier and the ground truth.

4.3.2 Chinese Data Test. To test the multilingual capabilities of our method, a Chinese classifier was implemented as defined in our methodology. The classifier performance was evaluated across the three testing sets mentioned in subsection 4.2. Figure 8 presents the results for classification of sarcasm in Chinese comments on all standard metrics. Similar as in English, the classifier achieves better scores for accuracy and precision. This behavior is sustained across different levels of agreement and different testing sets, which again is an indicator of a well-balanced classifier. The importance of these results is that the presented method learns directly from the data; no external human knowledge is added. Although the performance metrics do not present very high scores, they still provide a sense that the features evaluated in this work can indeed provide some clues for sarcastic posts.

5 CONCLUSIONS

This work brings focus to the importance of understanding the inherent relationship between sarcastic comments and emotions. Additionally, it highlights the value of platforms being used, how users interact on them, and, more importantly, how we can make use of these behaviors when working toward identifying sarcasm.

In this particular case, background knowledge of Facebook pages from news media gave some particular insights that later played an important role in the development of the method. Some examples include common behaviors of internet trolls, the usage of "reaction" buttons, and other commenting tendencies. There is still much to be done in terms of developing precise, efficient, and effective methods for sarcasm detection. To the best of our knowledge, this is the first work using Facebook reactions as emotion signals for any sentiment-related task. While this study does not address all the complexity of user emotions and reactions, nor does it try to achieve state-of-the-art performance, it does provide some evidence that the features evaluated in this work can be useful in the task of sarcasm classification. More importantly, we have helped to define a method where these features can be learned by the system without external human knowledge.

This work also provides a brief view on some behaviors related to the data and cultural and language differences. The extracted emotion patterns can serve as a summary of the data being addressed or as an embedded representation of the content. They can also intuitively reflect many characteristics of the audience based on the language or expressions they use. It is also interesting to notice the differences of emotion reactions in the data collection. Different languages reflect different uses of emotions in posts, as it was observed with the higher percentage of "Love" posts by English users than those in Chinese. Although it is still risky to call it a cultural difference, it can be assumed that there was a situational difference where at the time one of the language groups was leaning more toward those emotions given the ongoing events.

6 FUTURE WORK

Extensions of this work will attempt to improve the performance metrics to compete with the state of the art. In order to obtain a higher recall, for instance, more examples of sarcastic comments (explicit and not explicit) may be used in the training process for the patterns to learn their characteristics. More context can be included in the analysis, such as the elicited polarity of the original news post. Additionally, other data sources must be considered, since some of the behaviors are very particular to news sites; this might pose difficulty when trying to perform an extensive study. Data is already being collected in Spanish to develop a similar system and test if the method is indeed multi-lingual, or to evaluate to what degree it is. Additional studies that can be derived from this work include the study of cultural differences in sarcastic posting behavior. Likewise, evaluating which kinds of posts are more likely to receive sarcastic replies can also be carried out. Finally, it is possible to study the role of language in the usage, understanding, and proliferation of sarcasm.

REFERENCES

[1] Carlos Argueta, Fernando H Calderon, and Yi-Shin Chen. 2016. Multilingual emotion classifier using unsupervised pattern extraction from microblog data. *Intelligent Data Analysis* 20, 6 (2016), 1477–1502.

[2] David Bamman and Noah A Smith. 2015. Contextualized Sarcasm Detection on Twitter.. In *ICWSM*. Citeseer, AAAI, Oxford,UK, 574–577.

[3] Santosh Kumar Bharti, Korra Sathya Babu, and Sanjay Kumar Jena. 2015. Parsing-based sarcasm sentiment recognition in twitter data. In *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on.* IEEE, ACM, Paris, France, 1373–1380.

[4] Erik Cambria. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems* 31, 2 (2016), 102–107.

[5] Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd international conference on computational linguistics: posters.* Association for Computational Linguistics, 241–249.

[6] Rui Fan, Jichang Zhao, Yan Chen, and Ke Xu. 2014. Anger is more influential than joy: Sentiment correlation in Weibo. *PloS one* 9, 10 (2014), e110184.

[7] Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford* 1, 12 (2009).

[8] Roberto González-Ibánez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in Twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2.* Association for Computational Linguistics, 581–586.

[9] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. 2013. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web.* ACM, 607–618.

[10] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media.*

[11] Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing Context Incongruity for Sarcasm Detection.. In *ACL (2)*. 757–762.

[12] Christine Liebrecht, Florian Kunneman, and Antal van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets# not. *WASSA 2013* (2013), 29.

[13] Andrew Lipsman, Graham Mudd, Mike Rich, and Sean Bruich. 2012. The power of fiLikefi. *Journal of Advertising research* 52, 1 (2012), 40–52.

[14] Bing Liu. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions.* Cambridge University Press.

[15] Peng Liu, Wei Chen, Gaoyan Ou, Tengjiao Wang, Dongqing Yang, and Kai Lei. 2014. Sarcasm detection in social media based on imbalanced classification. In *International Conference on Web-Age Information Management.* Springer, 459–471.

[16] Edwin Lunando and Ayu Purwarianti. 2013. Indonesian social media sentiment analysis with sarcasm detection. In *Advanced Computer Science and Information Systems (ICACSIS), 2013 International Conference on.* IEEE, 195–198.

[17] Diana Maynard and Mark A Greenwood. 2014. Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis.. In *LREC*. 4238–4243.

[18] Christopher Mueller. 2016. Positive feedback loops: Sarcasm and the pseudo-argument in Reddit communities. *Working Papers in TESOL and Applied Linguistics* 16, 2 (2016), 84–97.

[19] Aminu Muhammad, Nirmalie Wiratunga, and Robert Lothian. 2016. Contextual sentiment analysis for social media genres. *Knowledge-Based Systems* 108 (2016), 92–101.

[20] Alvaro Ortigosa, José M Martín, and Rosa M Carro. 2014. Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior* 31 (2014), 527–541.

[21] Robert Plutchik and Henry Kellerman. 1980. *Emotion: theory, research and experience. Vol. 1, Theories of emotion.* Academic Press.

[22] Soujanya Poria, Erik Cambria, and Alexander F Gelbukh. 2015. Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-level Multimodal Sentiment Analysis.. In *EMNLP*. 2539–2544.

[23] Tomás Ptácek, Ivan Habernal, and Jun Hong. 2014. Sarcasm Detection on Czech and English Twitter.. In *COLING*. Citeseer, 213–223.

[24] Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. 2015. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining.* ACM, 97–106.

[25] Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering* 74 (2012), 1–12.

[26] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as Contrast between a Positive Sentiment and Negative Situation.. In *EMNLP*, Vol. 13. 704–714.

[27] Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media.. In *EMNLP*. 1815–1827.

[28] Yaowei Wang, Yanghui Rao, Xueying Zhan, Huijun Chen, Maoquan Luo, and Jian Yin. 2016. Sentiment and emotion classification over noisy labels. *Knowledge-Based Systems* 111 (2016), 207–216.

[29] Robert E Wilson, Samuel D Gosling, and Lindsay T Graham. 2012. A review of Facebook research in the social sciences. *Perspectives on psychological science* 7, 3 (2012), 203–220.

[30] Jichang Zhao, Li Dong, Junjie Wu, and Ke Xu. 2012. Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 1528–1531.