Detecting Public Sentiment of Medicine by Mining Twitter Data

Daisuke Kuroshima[1], Tina Tian[2]
Department of Computer Science, Manhattan College, New York, USA

***Abstract*—The paper presents a computational method that mines, processes and analyzes Twitter data for detecting public sentiment of medicine. Self-reported patient data are collected over a period of three months by mining the Twitter feed, resulting in more than 10,000 tweets used in the study. Machine learning algorithms are used for an automatic classification of the public sentiment on selected drugs. Various learning models are compared in the study. This work demonstrates a practical case of utilizing social media in identifying customer opinions and building a drug effectiveness detection system. Our model has been validated on a tweet dataset with a precision of 70.7%. In addition, the study examines the correlation between patient symptoms and their choices for medication.**

Keywords*—*Twitter; social media; data mining; public health

I. Introduction

Twitter, a microblogging service, has gained rapid popularity over the decade [1]. Massive quantities of real-time, fine grained microblog messages (also known as tweets) are available on Twitter. It is estimated that the social network attracts 321 million monthly active users worldwide, posting more than 500 million tweets everyday [2]. Published tweets can be accessed through Twitter's web portal or extracted programmatically using its Application Program Interface (API) [3].

Due to its rapid growth and the accessibility of the massive quantities of tweets, Twitter has become a valuable information resource for various applications. For example, enterprises have studied the usefulness of Twitter in organizational communication and information gathering [4]. Researchers have monitored real-time activities on Twitter to detect earthquakes [5]. Furthermore, Twitter has been used in studying political campaigns [6].

One particularly interesting research field is to apply social network data in the medical domain. Past research suggested that social media serve as valuable tools to involve patients more in their care and promote more effective communications between physicians and patients [7]. Currently, Twitter is the most popular platform of social media used for healthcare communications [8].

In this paper, we present a computational approach that collects, processes and analyzes Twitter data for detecting public sentiment of medicine. Machine learning models are used for an automatic classification, which is able to examine tweets that show positive or negative sentiment. The results from the supervised classification study demonstrate how Twitter can be utilized to identify patterns of customer opinions.

The rest of the paper is organized as follows. Section II reviews the related work. In Section III, we describe the data set used in the experiment and explain the methods and algorithms adapted in analyzing the data. Section IV presents the results of the study and Section V concludes the paper and proposes future directions.

II. Related Work

Social media, particularly Twitter, has gained increasing attention in medicine [9]. This is mainly due to the broad reach of Twitter users; anyone with a Twitter account is able to publish public tweets of up to 140 characters [10]. Compared to the traditional approach of evaluating drug efficacy where information comes from limited surveillance resources, Twitter greatly increases the number of people who can contribute to the discussion. Therefore, larger scale data can be leveraged for studying the effectiveness of medication.

Tweets can provide critical opinions and first-hand reviews on drugs based on patient experience. Communications on Twitter come from diverse backgrounds, making it a unique source of information gathering from the general population. Moreover, Twitter provides real time and direct surveillance. For example, Lee et al. used Twitter data for real-time disease surveillance on flu and cancer [11]. Gesualdo monitored tweets for allergic disease surveillance [12].

Past research has shown the predictive power of Twitter for health care. Tweets have been used to collect evidence about post-market pharmacovigilance [13] [14]. Aramaki et al. have used Twitter to detect influenza epidemics [15]. Baumgartner et al. utilized Twitter for discovering emergent online communities of cannabis users [16]. Recently, Twitter data were used as an information source to detect drug abuse in real time [17]. In addition, the study of Twitter updates successfully tracked the spread of cholera in Haiti [18]. Overall, Twitter has been proven to be a valuable knowledge source in tracking natural disasters, infectious disease outbreaks and drug use.

In this paper, we describe a computational method that identifies the public sentiment of over-the-counter (OTC) drugs using Twitter. Specifically, common pain relievers are targeted in the experiment. Self-reported patient data are collected by mining the Twitter feed. Moreover, the study shows the uses of the selected drugs among Twitter users, presenting a different perspective compared to the drug facts provided by pharmaceutical companies.

III. Data Set and Methodology

The system performs a sentiment analysis on a Twitter tweet corpus regarding drugs collected from January 2019 to April 2019. In this section, we discuss the data set used in the study and how we processed the tweets in order to determine their sentiment.

A. Data Set

In this work, a list of four OTC painkillers is examined. They are Advil, Aleve, Motrin and Tylenol. We name it List A. Brand names are chosen over names of the substances, as they are more frequently mentioned by Twitter users. Both Advil and Motrin contain ibuprofen. Aleve's main substance is naproxen while Tylenol majorly contains aspirin.

It is possible to collect a subset of Twitter feed by running a keyword search using Twitter's Search API. In order to extract more relevant tweets for learning the sentiment, we apply a second list of keywords, which includes symptoms of the patients taking drugs in List A, reported by pharmaceutical companies. Table I shows a sample of keywords included in the new list, named List B. This approach also enables us to study uses of the drugs from the general public.

Synonyms of the keywords are included in List B. For example, "throw up" and "puke" are added to the list as synonyms of "vomit". Different tenses are taken into account as well. For instance, besides "puke", the list also contains "puked" and "puking". Twitter is known for its informal language style [19]. Therefore, List B has included several casual words and phrases, such as "tummy", which is an informal way to describe the stomach.

Both List A and List B are passed to the Twitter Search API for extracting the dataset. A tweet is only selected if it contains at least one drug from List A and one keyword from List B. In this study, we target tweets that are written in English and published in the United States. During the three months of data collection, more than 10,000 tweets were extracted. Duplicated tweets, such as retweets, were removed. The remaining 6,447 distinct tweets form the final dataset for the study.

TABLE. I. Sample of Keywords in List B

stomach	cough	fever
headache	vomit	bloat
nausea	swelling	blood
itching	rash	sore throat
running nose	stuffy nose	sneezing
pain	belching	stiff
irritated eyes	lower back	difficulty sleeping
cry	swallow	muscle

TABLE. II. Breakdown of the Dataset

Drug	**Number of tweets**
Advil	2,162
Aleve	455
Motrin	502
Tylenol	3,680

Table II reveals a breakdown of the dataset concerning each specific drug. The total number of tweets in the table slightly exceeds the size of the dataset, as some tweets include more than one drug in List A. As one can see, Advil and Tylenol are more popular choices among Twitter users, while Tylenol being the most frequently mentioned drug in the experiment.

B. Data Pre-Processing

Before we deliver the dataset to machine learning methods for sentiment detection, it is necessary to pre-process the raw data. As seen in Fig. 1, the process includes data manipulations, such as data cleaning, data splitting and text vectorization to properly prepare the training set. The training data are then labeled with sentiment tags and passed to machine learning algorithms for further studying the test dataset. The rest of this section elaborates the processing procedure.

The first stage of data pre-processing is text cleaning. Extracted tweets are converted to all lowercase letters. URLs in tweets are removed, as they do not contribute to sentiment detection. Tweets with selected user mentions are eliminated from the study. One example is the user mention @Advil, which appears in our dataset as Advil is one of the search keywords. However, username @Advil belongs to a personal Twitter account, who has no relation with Advil the drug.

Next, we split the dataset into training set and test set to be used with machine learning algorithms. Among all the tweets, 259 of them are randomly selected to form the training set. The test set consists of the remaining 6,188 tweets. Each tweet in the training set is manually classified as positive sentiment or negative sentiment. They are labeled P or N respectively. A tweet is evaluated as positive if the Twitter user (patient) published a positive experience with the drug. On the other hand, a tweet is labeled negative if the user reported the drug being ineffective. As a result, 133 tweets in the training set are classified as positive and the remaining 126 tweets are labeled negative.

The training dataset and the sentiment labels are stored in an .arff file in order to be processed by machine learning tools. Fig. 2 shows a sample of the training data file. As one can see, two attributes are included in the relation. One attribute is the tweet, which is a string type, and the other is its classification, which comes from set {P, N}. The rest of the file consists of the data section, in which each line contains an example with the two attributes separated by a comma.

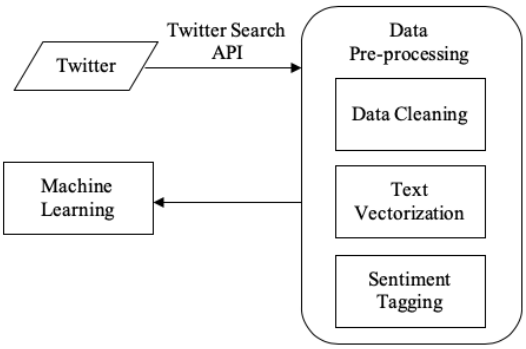

Fig. 1. System Architectural Overview.

@relation 'training.arff'

@attribute tweet string
@attribute classification {P, N}

@data
"i'd be sick because i'm allergic to advil", N
"when i say i was on a million meds, i mean it. i was taking a steroid pack to help me heal, melatonin to sleep, advil to help my pain and i had anti nausea meds like i was taking easily like 10 pills a day for a good month oop", P
"@sar_free well take some tylenol so that fever can go away", P
"i have such a headache i took tylenol to stop this exact thing", N

Fig. 2. Sample of Training Data.

In order for our machine learning models to learn the correlation between a tweet and its sentiment classification, text vectorization is used to parse the data. In the process, tweet strings are converted to word vectors, as known as bag of words. Tweets are tokenized by applying pre-defined delimiters, such as blank spaces and punctuation signs. For example, tweet "i'd be sick because i'm allergic to advil" is converted to vector [i, d, be, sick, because, i, m, allergic, to, advil]. A list of stop words is applied to eliminate neutral and nonsemantic terms in a vector, such as the, in, of, etc. Additionally, stemming is used to words with the same root. For example, cat, cats, catlike and catty are considered the same term.

Overfitting is a common issue in machine learning. To tackle the problem, we select the top 100 most frequent words from each classification. That is, 100 terms from the positive tweet corpus and 100 terms from tweets with the negative annotation. Together they form the vocabulary of this study. In this work, the most frequent words are determined by calculating their term frequency-inverse document frequency (tf-idf). The following formulas describe how tf-idf of a term is computed.

Let *t* be a term and *d* be the tweet that contains term *t*. The term frequency of *t*, denoted as *tf(t,d)*, can be calculated as

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \quad (1)$$

where $f_{t,d}$ stands for the number of appearances of term *t* in the tweet and $\sum_{t' \in d} f_{t',d}$ represents the total number of occurrences of all words in tweet *d*. Let *D* be the entire document, specifically, the collection of tweets in our experiment. The inverse document frequency of term *t*, symbolized as *idf(t,D)*, can be computed using the following formula.

$$idf(t,D) = log \frac{N}{|\{t' \in D : t \in d\}|} \quad (2)$$

where *N* stands for the total number of tweets in the dataset. $\{t' \in D : t \in d\}$ represents the set of tweets in our data collection that contain term *t*. For example, there are 6,447 tweets in the dataset and the word sick appears in 189 of them. Thus, the inverse document frequency of term sick is computed as log(6447/189), which approximates 5.1. The final term frequency-inverse document frequency of term *t* is the product of the two previously calculated frequencies, as shown in the formula below.

$$tfidf(t,d,D) = tf(t,d) \cdot idf(t,D) \quad (3)$$

Term frequency-inverse document frequency is used as the measure to select the most popular terms to be studied by machine learning models. As a result, a combined list of 141 top words from both the positive and the negative classifications constructs the vocabulary. The total size is less than 100 times 2, as some words overlap in the two classes.

IV. Results

As mentioned in Section III, our dataset consists of 6,447 distinct tweets that contain keywords from both a list of medicines and a list of symptoms. The coverage of symptoms in the user-reported data enables us to further study the uses of the drugs. Overall in our experiment, headache is the most common cause to seek over-the-counter pain relievers. Other frequently mentioned symptoms are ear pain and back pain.

Fig. 3 shows a bar chart of the top five most common uses for each medicine. A bar in the graph represents the percentage of a particular symptom reported for a drug. Common symptoms include headache, fever, ear pain, back pain, etc. The term general pain is used to represent unspecified pain reported by the patients. For example, our study shows that the main uses of Advil are headache, general pain and back pain. Among them, headache is the dominant cause of Advil consumption, reported by 33% of the Advil users.

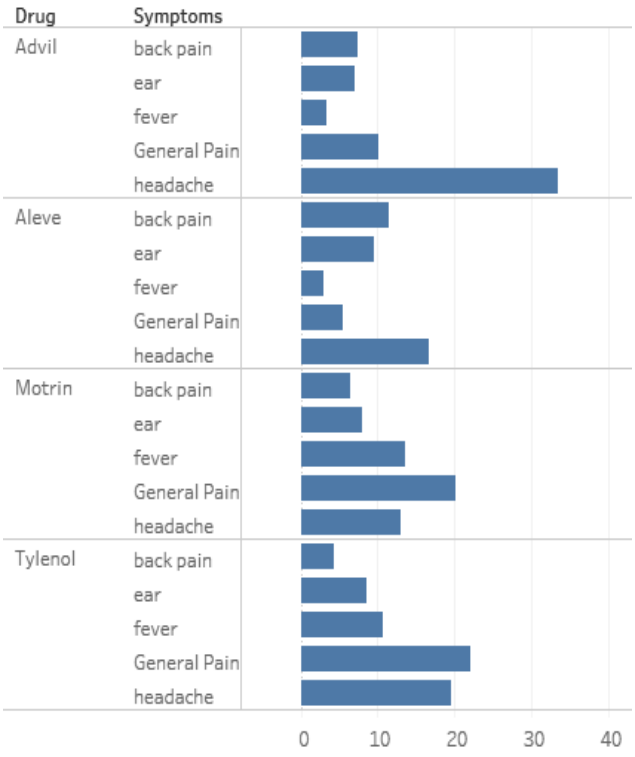

Fig. 3. Use of Each Drug in Percentage.

Fig. 4 reveals the co-occurrence patterns between drugs in our study and their reported uses. Circles are color coded in order to distinguish among different drugs. For example, Motrin consumption is symbolized with color red and patient cases of Advil are annotated with blue circles in the figure. Each circle represents the tweets in our dataset that reported a particular use of a drug. For example, circle "Advil headache" in the figure symbolizes the set of tweets that contain both keywords advil and headache. The size of a circle is in proportion to the number of tweets it represents. Thus, large circles are a sign of frequent co-occurrences between a drug and its reported use.

As seen in Fig. 4, patients with headache tend to seek Advil. In our study, it is found that people with the symptom of fever prefer Tylenol and Motrin over others. An interesting observation is that a large portion of Aleve consumers use it for ear pain. The dataset also suggests that Tylenol is a popular choice for other specific pain, such as hip pain, backache, muscle cramp and chronic pain. Moreover, 50% of the patients who reported to show symptoms of sore throat and swelling chose Aleve over other pain relievers in the study.

In order to render the most accurate result in examining the sentiment of tweets, several machine learning models are applied in this work. They are decision tree learning, random forest, support vector machine, naïve Bayes and k-nearest neighbors. The study compares their accuracy and selects the algorithm with the highest precision for automatic sentiment classification. As mentioned in Section III, the training set consists of 259 randomly selected tweets in the dataset. Each tweet is labeled with a sentiment of positive or negative. In order to compare different machine learning algorithms, the training data are divided into two sets. One set contains 90% of the training data, which is used to train the different machine learning models. The remaining 10% of the training data construct the second set, which is used to validate the classification results. In this work, 10-fold cross validation is applied to achieve better precision.

Every machine learning model generates its predicted result regarding each tweet in the validation set. The prediction is a classification of positive or negative sentiment. If a tweet in the validation set is labeled positive and the learning model successfully predicts it, we mark it as a case of True Positive (TP). On the other hand, if the model delivers a positive classification while the tweet is annotated negative, we record the instance as False Positive (FP). Similarly, a tweet is considered True Negative (TN), if the learning algorithm correctly reports a negative sentiment. Otherwise, if the model fails to predict a positive label, the case is noted as False Negative (FN). The four categories summarize all possible scenarios from the validation process. An example of each category is shown in Table III.

To evaluate the accuracy of each machine learning model, three measures are considered in this study. They are precision, recall and F-measure. All the three parameters can be computed using the number of instances included in the four categories viewed in Table III. The following formulas show the calculation of the three accuracy measures, where TP, FP, TN and FN represent the number of tweets belonged to each scenario.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F\text{-}measure = \frac{Presision \cdot Recall}{Precision+Recall} \quad (6)$$

Table IV reveals the precision, recall and f-measure from validating selected machine learning models. Each model is associated with two sets of accuracy measures. One set is based on the positive classification and the other is built on the negative classification. Finally, their weighted average is calculated, which shows in bold in the table.

As seen in Table IV, among all the learning models, naïve Bayes provides the highest average precision, recall and f-measure. Therefore, it is chosen as the algorithm to determine the sentiment classifications of the test data. As a result, 3,068 tweets in the test set are categorized with positive sentiment and the rest 3,377 tweets are classified as negative.

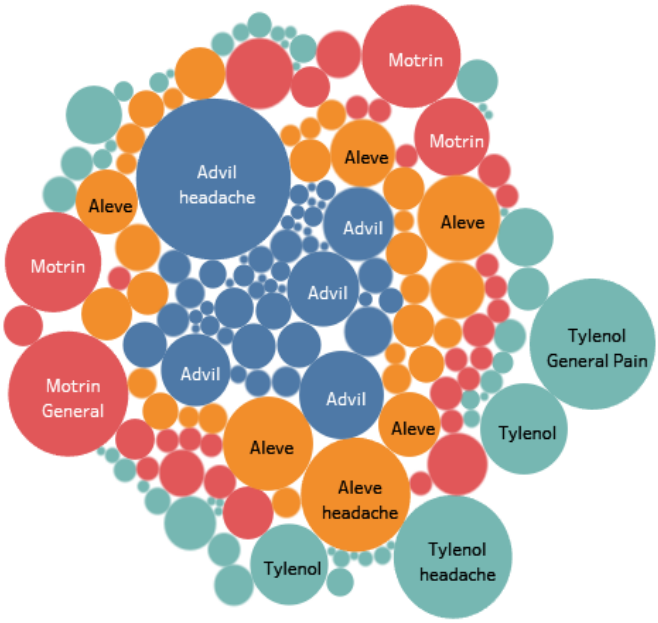

Fig. 4. Co-Occurrence Patterns between Drugs and their uses.

TABLE. III. EXAMPLES OF CLASSIFICATION RESULTS

Category	Example
TP	Mix salt in warm water and wash around your mouth. It alleviates the pain a little. Advil liquid gels do work too lol
FP	Unfortunately I DO have stomach issues, after playing that game for about ten years. :(I used to use extra-strength advil AND extra-strength Tylenol to knock the pain out, but it caught up to me this year.
TN	The IUD is gone and the tylenol is taken and STILL THE PAIN.
FN	I have a headache today... should take an Advil

TABLE. IV. PRECISION, RECALL AND F-MEASURE

Learning Model	Classification	Precision	Recall	F-Measure
Decision Tree	Positive	57.8%	65.5%	61.4%
	Negative	64.9%	57.1%	60.8%
	Weighted Average	**61.5%**	**61.1%**	**61.1%**
Random Forest	Positive	64.3%	55.8%	59.7%
	Negative	64.5%	72.2%	68.2%
	Weighted Average	**64.4%**	**64.4%**	**64.2%**
Support Vector Machine	Positive	69.9%	57.5%	63.1%
	Negative	67.1%	77.8%	72.1%
	Weighted Average	**68.4%**	**68.2%**	**67.8%**
K-Nearest Neighbors	Positive	62.6%	68.1%	65.3%
	Negative	69.0%	63.5%	66.1%
	Weighted Average	**66.0%**	**65.7%**	**65.7%**
Naïve Bayes	Positive	69.7%	67.3%	68.5%
	Negative	71.5%	73.8%	72.7%
	Weighted Average	**70.7%**	**70.7%**	**70.7%**

V. CONCLUSIONS AND FUTURE WORK

In this work, we built a drug sentiment classification system based on Twitter data. The system is able to automatically identify patients' opinions on selected drugs. The study demonstrated that public sentiment of medicine can be detected using data on social media, such as tweets. Our model has been validated on a real-world dataset with a precision of 70.7%. Additionally, the study investigated the correlation between patient symptoms and their choices for medication.

In the future, it is planned to apply the same methodology on prescription drugs and newly released medications. Learning public sentiment of new drugs can be particularly crucial, providing valuable feedback for patients and healthcare providers. Moreover, future work involves expanding the dataset, which will include a larger pool of tweets for training the learning models. It is possible for the system to achieve a higher precision rate with an expanded set of training data.

Currently, this work only considers binary classifications. Thus, the sentiment of a tweet is labeled as either positive or negative. However, it is observed that some tweets do not indicate an identifiable opinion regarding the drugs in the study. In other words, they have a neutral sentiment. These tweets have been eliminated from the training set. Nevertheless, it is possible that the test data may contain a subset of neutral tweets that were mislabeled as positive or negative by our learning model. In the future, it is planned to adapt the one-vs-all classification method to tackle neutral sentiment.

REFERENCES

[1] C. C. Aggarwal, Social Network Data Analytics. Springer, 2011.

[2] Twitter, http://twitter.com, retrieved 09/01/2019.

[3] Twitter Search API, https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html, retrieved 09/01/2019.

[4] A. Archambault and J. Grudin, "A longitudinal study of Facebook, LinkedIn, & Twitter use," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2741-2750, 2012.

[5] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors," Proceedings of the 19th International Conference on World Wide Web, pp. 851-860, 2010.

[6] I. Vegas, T. Tian, and W. Xiong, "Charactering the 2016 U.S. presidential campaign using Twitter data,", Journal of Advanced Computer Science and Applications, vol. 7, no. 10, 2016.

[7] H. Cayton, "The alienating language of health care," Journal of Royal Society of Medicine, vol. 99, no. 10, pp. 484 -484, 2006.

[8] W. Raghupathi and V, Raghupathi, "Big data analytics in healthcare: promise and potential," Health Information Science and Systems, vol. 2, no. 1, 2014.

[9] K. Chretien, J. Azar, and T. Kind, "Physicians on Twitter," Journal of the American Medical Association, vol. 305, no. 6, pp.566 -568, 2011.

[10] Twitter,https://developer.twitter.com/en/docs/basics/counting-characters.html, retrieved 09/01/2019.

[11] K. Lee, A.A. Agrawal, and A. Choudhary, "Real-time disease surveillance using Twitter data: demonstration on flu and cancer," Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1474 – 1477, 2013.

[12] F. Gesualdo, G. Stilo, A. D'Ambrosio, E. Carioni, E. Pandolfi, P. Velardi, A. Fiocchi, and A.E. Tozzi, "Can Twitter be a source of information on allergy? correlation of pollen counts with tweets reporting symptoms of allergic rhino conjunctivitis and names of antihistamine drugs," PLoS One, vol. 10, no. 7, 2015.

[13] N. Alvaro, M. Conway, S. Doan, C. Lofi, J. Overington, and N. Collier, "Crowdsourcing Twitter annotations to identify first-hand experiences of prescription drug use," Journal of Biomedical Informatics, vol. 58, pp. 280 - 287, December 2015.

[14] A. MacKinlay, H. Aamer, and A. Yepes, "Detection of adverse drug reactions using medical named entities on Twitter," Annual Symposium Proceedings, AMIA Symposium, pp. 1215 - 1224, 2017.

[15] E. Aramaki, S. Maskawa, and M. Morita, "Twitter catches the flu: detecting influenza epidemics using Twitter," Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1568 – 1576, 2011.

[16] P. Baumgartner and N. Peiper, "Utilizing big data and Twitter to discover emergent online communities of cannabis users," Substance Abuse: Research and Treatment, vol. 11, 2017.

[17] N. Phan, S. A. Chun, M. Bhole, and J. Geller, "Enabling real-time drug abuse detection in tweets," Proceedings of the 33rd International Conference on Data Engineering, 2017.

[18] R. Chunara, J.R. Andrews, and J.S. Brownstein, "Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak," The American Journal of Tropical Medicine and Hygiene, vol. 86, issue 1, pp. 39 - 45, 2012.

[19] K.L. Liu, W.J. Li and M. Guo, "Emoticon smoothed language models for Twitter sentiment analysis," Proceedings of the 26th AAAI Conference on Artificial Intelligence, pp. 1678-1684. 2012.