Opinion Mining and Sentiment Polarity on Twitter and Correlation Between Events and Sentiment

Peiman Barnaghi[1,2] and John G. Breslin[1]
[1]Insight Centre for Data Analytics,
National University of Ireland, Galway
IDA Business Park, Lower Dangan,
Galway, Ireland
peiman.barnaghi@insight-centre.org
john.breslin@nuigalway.ie

Parsa Ghaffari[2]
[2]AYLIEN Ltd.
4th Floor, Equity House
16/17 Ormond Quay Upper, Dublin7
Dublin, Ireland
parsa@aylien.com

Abstract—**Twitter, as a social media is a very popular way of expressing opinions and interacting with other people in the online world. When taken in aggregation tweets can provide a reflection of public sentiment towards events. In this paper, we provide a positive or negative sentiment on Twitter posts using a well-known machine learning method for text categorization. In addition, we use manually labeled (positive/negative) tweets to build a trained method to accomplish a task. The task is looking for a correlation between twitter sentiment and events that have occurred. The trained model is based on the Bayesian Logistic Regression (BLR) classification method. We used external lexicons to detect subjective or objective tweets, added Unigram and Bigram features and used TF-IDF (Term Frequency-Inverse Document Frequency) to filter out the features. Using the FIFA World Cup 2014 as our case study, we used Twitter Streaming API and some of the official world cup hashtags to mine, filter and process tweets, in order to analyze the reflection of public sentiment towards unexpected events. The same approach, can be used as a basis for predicting future events.**

Keywords—Sentiment Analysis, Opinion Mining, Stream Data Analysis, Polarity Detection, Sentiment Classification, Keyword Correlation, Natural Language Processing, Sentiment Mining, Twitter

I. INTRODUCTION

Twitter, one of the most common online social media and micro-blogging services, is a very popular method for expressing opinions and interacting with other people in the online world. Twitter messages provide real raw data in the format of short texts that express opinions, ideas and events captured in the moment. Tweets (Twitter posts) are well-suited sources of streaming data for opinion mining and sentiment polarity detection [1]. Opinions, evaluations, emotions and speculations often reflect the states of individuals; they consist of opinionated data expressed in a language composed of subjective expressions [2]. In this paper, we examine the effectiveness of a commonly used text categorization method called Bayesian Logistic Regression (BLR) Classification for providing positive or negative sentiment on tweets. We use extracted Twitter sentiment to look for correlations between this sentiment and major FIFA World Cup 2014 events as our case study. The rest of the paper is organized as follows: In section 2, we discuss sentiment analysis on corpora, section 3, gives details about the proposed model, data and pre-processing methods, section 4, discusses our trained model. In section 5, we present our case study for the correlation task we give details of our feature based approach. Section 6, gives details about two main machine learning methods for sentiment classification and evaluations. Section 7, discusses correlation between events and sentiment. We conclude and give future directions of research in section 8.

II. RELATED WORK

A broad overview of some of the machine learning techniques used in sentiment classification is provided by Pang et al. [3]. They provided an overview of three well-known machine-learning methods for text categorization, including Naïve Bayes, Logistic Regression and Support Vector Machine. They used movie reviews for classifying sentiment as positive or negative. Opinions are classified as one of two opposing sentiment polarities (positive or negative), however, they may also be labeled as neutral when there is a lack of opinion in the text or the opinion is located in between these two polarities. This kind of labeling can be used to summarize the content of opinionated texts and documents. A wide variety of features can be necessary for opinion and polarity recognition [1]. Many advanced methods and algorithms have been developed for text categorization during the last three decades [4-7] .The bag-of-words method is a standard approach and the most popular model for text categorization [8] as the concept is easy to understand and also helps improve performance [9, 10]. The bag-of-words method uses a vector of words in Euclidean space for representing the document where each word is independent from others and used as a feature for training a sentiment model [8].

Feldman et al. [11] discuss three levels of sentiment analysis: document, sentence and aspect-based. Sentiment at the document and sentence levels works well when the corpus refers to a single entity, but sentiment at the aspect level is fine-grained analysis when there are many aspects or attributes in a corpus with different opinions about each of them. An external lexicon (SentiWordNet) is used to support opinion mining and sentiment classification [12]. Pak et al [13] discuss tweet gathering methods for use in sentiment analysis. They use a specific lexicon of emoticons to reduce manual tweet tagging for sentiment classification. Based on happy and sad emoticons, the training set was split into positive and negative samples.

III. TWEET SENTIMENT ANALYSIS METHOD

A sentiment analysis model on Twitter data is shown in Fig. 1. It shows the different steps of pre-processing, feature extraction and filtering to train a model for polarity detection. Tweets contained useless information that the workflow is designed to order and clean using tokenization, uppercase conversion, stop-words filtering, stemming and lemmatization and also converting the content of messages such as username, URLs to general tags and hashtag detection to mark topics and keywords. As a final target, the trained model is used for finding correlations between tweets and major events in the World Cup.

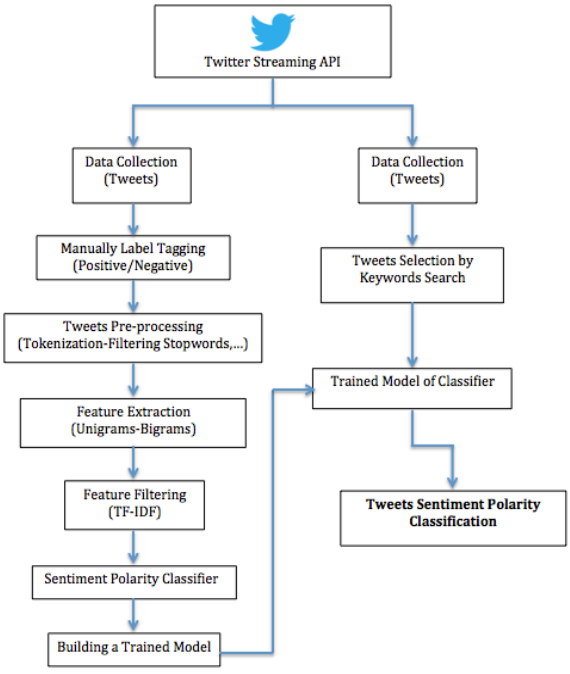

Fig. 1. Building a trained model for polarity detection and applying it to some tweets

A. Case Study of Correlation for Sentiment Analysis

After 64 football matches, 672 million tweets were posted, which were related to the 2014 World Cup [14]. In this paper, we use extracted twitter sentiment to look for correlations between this sentiment and major FIFA World Cup 2014 events. We use Twitter's Streaming API for mining tweets and processing them by filtering using some of the official World Cup hashtags (e.g. "#worldcup" and "#brazil2014"). This paper looks at some of the major talking points from the tournament that were extracted from Twitter data during the 2014 World Cup. The following events were two of the most incidents: Firstly, a Uruguayan football player was accused of biting an Italian player. Secondly, the elimination of Brazil during the tournament (regarded as one the best football teams in the world, and also the host country of the 2014 World Cup). These two events were followed by a huge number of positive and negative tweets with changing sentiment based on events and timestamps.

B. Tweet Collection

Data gathering was made up of two steps using Twitter's Streaming API: the first was collecting the data to use as a training set to build the model. This consisted of 4162 tweets manually labeled "positive" or "negative". The second step was collecting tweets during the World Cup tournament and processing them by filtering some of the official World Cup hashtags (e.g. "#WorldCup" and "#Brazil2014"), as well as team code hashtags (e.g. "#ARG" and "#GER"). In addition, the Twitter usernames of teams and players were used to extract tweets relating to events (e.g. "#suarez" and "#brazil") that occurred during the tournament. The data was in JSON format as a set of documents, one for each tweet [15].

IV. TWEET TEXT PRE-PROCESSING

As a first step towards finding a tweet's sentiment and in order to obtain accurate sentiment classification, we needed to filter out noise and meaningless symbols that do not contribute to a tweet's sentiment from the original text. All of the following steps had to be performed sequentially for all the tweets, in order to use them for training a model:

A. Tokenization

Tokenization is the process of splitting up a string into a list of tokens and constructing a bag-of-words and is the first step of pre-processing. It involves splitting the text with white spaces to form a list of individual words in each text. A word is a token in a sentence that can be used as a feature to train a sentiment classifier.

B. Removing Stop-Words

Stop-words such as articles, prepositions and short function words carry a connecting function in the sentence and have a high frequency of occurrence in the text. They can be removed from a bag-of-words since they do not affect the final sentiment of the text. This can be done by checking each word from the text against a dictionary (WEKA machine learning package is used) including stop words such as "and", "or", "still", "also", "able", "the", "as", "which" etc. and removing all the matching ones.

C. Twitter Symbols

There are some symbols which may be used in tweets; for example the word following after the "@" symbol is a username and "#" is used to mark topics or keywords in a tweet. All usernames and URLs were converted to generic tags (e.g. all @usernames tagged as "username"), and some mentions can be used to improve the performance of the sentiment classifier [16].

D. Stemming

Stemming is a technique used to remove affixes from a word replacing them with their roots reducing different forms of a word such as nouns, verbs, adjectives etc. to a common base form. (e.g. the words "analysis", "analyzed", "analyzing" and all other types of this word are converted to "analysi" after stemming). We used a WEKA package including SnowballStemmer and LovinsStemmer to perform the stemming operation. It helps to reduce the dimensionality of the bag-of-words and improves the output of sentiment classification [17].

V. BUILDING A TRAINED CLASSIFIER

Labeling an opinionated text and categorizing it overall into a positive or negative class is called sentiment polarity classification. The neutral label is used for more objective items that have a lack of opinion in the text, or where there is a mixture of positive and negative opinions therein [1]. We need to use all the subjective tweets, including positive or negative sentiment. There are methods of extracting the useful words in order to detect the sentiment of tweets. The following section discusses feature extraction and selection methods and external lexicons including positive and negative words to compare the extracted features with available pre-defined ones.

A. Feature Extraction

Selecting a useful list of words as features of a text and removing a large number of words that do not contribute to the text's sentiment is defined as feature extraction. It helps us to filter noise from the text and obtain a more accurate sentiment for a tweet.

1) Unigram features:

Unigrams are the simplest method of feature extraction and are defined as looking at one word at a time in a text, which can be extended to an N-gram in order to exploit the ordering of words. It can be used in different states of text such as characters, words or sentences.

2) N-gram features:

An N-gram feature is defined as taking a set of sequential words in a text; for example if N=2, it means looking at a pair of sequential words at a time, which is called a bigram.
Some related works based on unigrams show that the kind of dataset has an impact on classification performance. Pang et al. [3] show that unigrams yield better performance on movie reviews for sentiment polarity classification. As tweets are very short texts with a maximum length of 140 characters and most tweets are around 30 characters long, N-gram features with N=1 to 2 are used, which uses a sensible list of sequential words for sentiment classifiers.

3) External Lexicon:

Using external lexicons helps to improve the performance of a sentiment classifier. One of the common uses of external lexicon for sentiment includes a list of words with predefined positive or negative sentiment. We used some of the open sourced lexicons such as MPQA and SentiStrength to detect positive or negative sentiments based on available words in tweets and for prior polarity based on the degree of sentiment in a word. [16].

B. Feature Filtering

As discussed before, the size of the corpora means that a large number of features are retained, which forces us to use methods to select the top features to use in training the classifier. Term Frequency-Inverse Document Frequency (TF-IDF) is a numerical statistical method to filter the features by weighting and scoring each of the unigrams and N-grams using the frequency of words in the text [3].

VI. SENTIMENT CLASSIFIER

The tweet polarity classifier is trained based on N-grams features (N=1 to 2) using WEKA[1] as a machine learning framework. Cross validation as a repeated holdout method is used on the dataset by splitting it into 10 sections. This method selects 90% for the training set, and 10% for the testing set, repeating it on 10 different sections of dataset. Finally, the result is averaged over the rotated divided sections. The goal of using this method is to test the model in the training phase [18]. The following steps were taken for machine learning classification: 1. Pre-processing/cleaning the data; 2. Features generation; 3. Features selection; 4. Training the model and validation.

A. Bayesian Logistic Regression

The Bayesian Logistic Regression (BLR) model simultaneously selects features and provides shrinkage for performing text categorization. It uses a Laplace prior to avoid over-fitting and produces sparse predictive models for text data [19]. The Logistic Regression estimation of P(c|f) has the parametric form:

$$P(c|f) = \frac{1}{z(f)} \exp((\sum_{i} \lambda_{i,c} F_{i,c}(f, c))$$

Where z(f) is a normalization function, λ is a vector of weight parameters for the feature set [20], and $F_{i,c}$ is a binary

[1]Weka is a collection of machine learning algorithms for data mining tasks. (http://www.cs.waikato.ac.nz/ml/weka/)

function that takes as inputs a feature and a class label. It is defined as:

$$F_{i,c}(f, c^{\cdot}) = \begin{cases} 1, & n(f) > 0 \; and \; c^{\cdot} = c \\ 0, & otherwise \end{cases}$$

This binary function is triggered when a certain feature (unigram, bigram, etc.) exists and the sentiment is hypothesized in a certain way. For example, a feature function might eliminate if the bigram "still like" appears and the sentiment of the document is hypothesized to be positive [3].

B. Naïve Bayes

This method is a simple probabilistic classifier with a strong conditional independence assumption that it is optimal for classifying classes with highly dependent features. Adherence to one of the positive, neutral or negative classes is calculated for each tweet using the probability based on the Bayes theorem. Even though this method as a simple probabilistic classifier with a strong conditional independence [2] assumption has yielded acceptable results [13] it is not good enough in comparison with some other classifiers as outlined in this section. In Bayes' theorem, $P(C_i|E)$ is the probability that text document E is of class C_i and defines it as follows [21].

$$P(C_i|E) = \frac{P(C_i)P(E|C_i)}{P(E)} \quad C_i \in C$$

C. Trained Clasiifier Evaluation

The first phase of this work is an evaluation of how the BLR classifier affects the performance of a simple two-class (positive / negative) sentiment analyzer. The following table displays the corresponding values for each experiment.

Table 1. Tweet polarity classifiers based on N-gram features.

Experimental Results \ ML Methods		**BLR** (Pos)	**BLR** (Neg)	**NB** (Pos)	**NB** (Neg)
Correctly classified instances %	Uni-grams	**71.35**		66.21	
	Bigrams	**67.44**		63.62	
	Unigrams & Bigrams	**74.84**		66.24	
Precision	Unigrams	71.4	71.3	70	63.7
	Bigrams	67.1	67.8	64.8	62.6
	Unigrams & Bigrams	74.5	75.1	71.5	63.1
Recall	Unigrams	71.7	72	56.2	76.1
	Bigrams	67.7	67.1	58.9	68.3
	Unigrams & Bigrams	75.1	74.6	53.5	78.9
F- score	Unigrams	71.1	71.6	62.3	69.4
	Bigrams	67.4	67.5	61.7	65.4
	Unigrams & Bigrams	74.8	74.9	61.2	70.1

VII. SENTIMENT ANALYSIS BASED ON WORLD CUP 2014 EVENTS

Tweets can provide a reflection of public sentiment when taken in aggregation during special events such as the FIFA World Cup. In this paper, sentiment analysis was carried out using our trained model for some of the major events that occurred during the tournament [22]. We analyzed only English tweets from the 30 million gathered tweets because a language analysis of the World Cup tweets showed that the 51.56% of tweets were in English. The positive, negative or neutral polarity values of these tweets were used to see what these values are for different entities and how they change over time, as a result of various events.

A. Correlation Between Event and Sentiment

To find the correlation between sentiment and events, we used a timestamp to associate each tweet and occurred events during the tournament. We computed the correlation using the Pearson correlation coefficient comparing two normalized time series of sentiment polarity and occurred events scores.

$$r_{xy} = \frac{\sum_{i=0}^{n}(x_i - x)(y_i - y)}{\sqrt{\sum_{i=0}^{n}(x_i - x)^2(y_i - y)^2}}$$

To calculate the sentiment score, if there were no negative sentiments, the ratio would be 1 and if there were more negative sentiments, the ratio would be closer to -1 [16].

1) A Major Event During The World Cup

We tried to extract all related tweets for a major event and to uncover correlations between the tweets and the event that occurred. During the FIFA World Cup 2014, on June 24th, an Uruguayan player, Luis Suarez was accused of biting an Italian defender, Giorgio Chiellini. The event was followed by a large volume of negative tweets on Twitter. Using the trained model, sentiment classification was performed on all tweets that mentioned the player's name. The sentiment classification output (Figure 2) shows that the trend of tweet polarity is divided into three different parts of sentiment for the aforementioned player. The first part consists of the polarity values of all tweets before the biting incident. There is a fluctuation of sentiment polarity rates based on player performance and match results. Almost all of these sentiments are positive with different strengths (such as strongly positive) or else neutral. The second part of the sentiment polarity shows the beginning of a negative trend after the incident. Almost all tweets are negative with different rates. The third part of the sentiment polarity starts when Suarez issued an apology on June 30th, which seems to have been

satisfactory for the Twitter community and a positive trend starts growing and reaches a peak level of positive polarity when he signed his new contract with Barcelona FC. Fig. 2 shows the sentiment polarity based on the following sections and details:

(1) Suarez allegedly bites Italy's defender
(2) Suarez issues an apology
(3) Suarez signed a contract with Barcelona

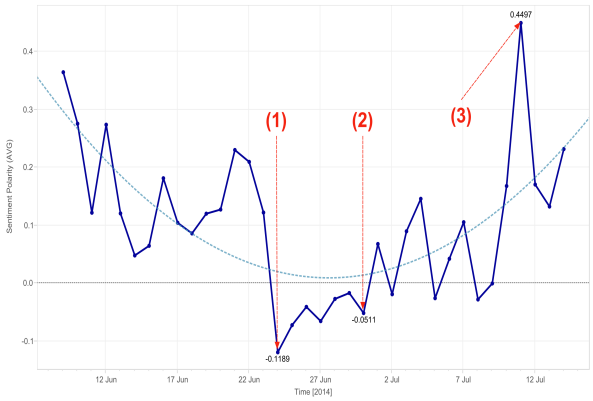

Fig. 2. Sentiment polarity of tweets about Luis Suarez during the world cup 2014

2) Elimination of Brazil as Another Major Event

Another major event during the tournament was the elimination of Brazil (regarded as one the best football teams in the world and also the host country team of the 2014 World Cup). Fig. 3. shows the trend with the average sentiment polarity based on positive or negative tweets and overall polarity (middle diagram). There are changes to the polarity of sentiments/tweets after each winning or losing match.

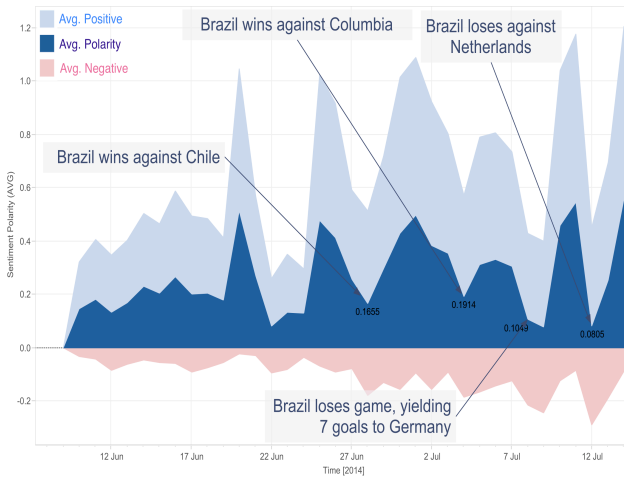

Fig. 3. Average sentiment polarity of tweets for the Brazilian team.

After losing the match against Germany (1 - 7) and before the 3rd place playoff game between Brazil and Netherlands, there is a significant change of polarity of tweets from negative to positive. It shows that people were still hopeful that the previous result might bring out the best in the Brazilians last game, but the direction changed again after conceding three goals against the Netherlands during 3rd place plya-off. Fig. 4. shows the increase in the number of tweets posted during the match and figure 5 shows the polarity changes of sentiments for Brazil's last match.

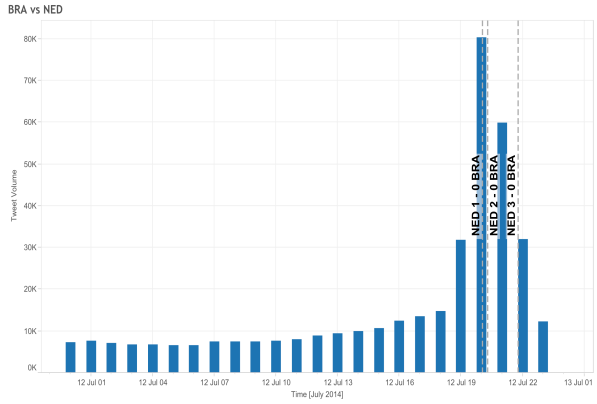

Figure 4. Tweet volume during the match

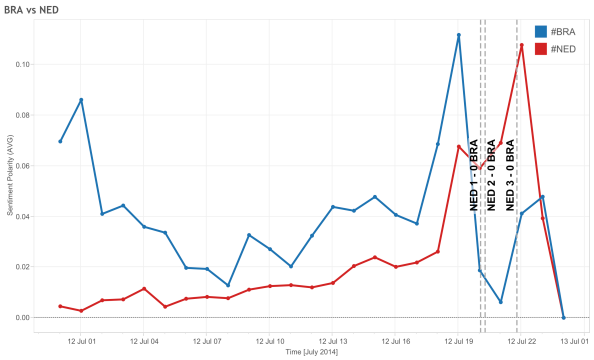

Figure 5. Average sentiment polarity during the Brazil and Netherlands match

VIII. FUTURE WORK

Future work will focus on the polarity classification of scalable topic-level streaming feeds, with classification of a streaming feeds' sentiment towards a given topic (and not just a keyword). The next step can be defined as; trend detection relating to a topic on a set of streaming feeds, to determine the polarity of the target topics. Also, determining the degree of polarity can be used to show the sentiment strength (such as strongly positive/negative or weakly positive/negative or neutral).

IX. Conclusions

We can use average sentiment polarity measures for various entities and events to see how positively or negatively people react or talk about them. Analyzing the sentiment of tweets gives an interesting insight into the opinions of the public in relation to a certain event. Analyzing Twitter posts allows the extraction of detailed insights into opinions and trends around sporting events such as the FIFA World Cup, players, teams, etc. and how they change over time during a critical event or after unethical behavior. In this paper, a sentiment classification model was trained based on Twitter data using text features. We extracted sentiment polarity for some major events that occurred during the World Cup using our trained model. The experimental results show the positive and negative reaction of people towards such events and how it can change based on incidents during those events. This kind of sentiment analysis helps us to use Twitter data for extracting patterns based on opinionated texts. In addition, teams, players, etc. can receive an overall sentiment in relation to their performance and behavior that could be used to help to improve the quality of matches by highlighting controversial ethical issues as well.

Acknowledgment

This project has emanated from research conducted with the financial support of the Irish Research Council (IRC) under Grant Number EBPPG/2014/30 and with Aylien Ltd. as Enterprise Partner. This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight).

References

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and trends in information retrieval,* vol. 2, pp. 1-135, 2008.

[2] A. Bifet and E. Frank, "Sentiment knowledge discovery in twitter streaming data," in *Discovery Science*, 2010, pp. 1-15.

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 2002, pp. 79-86.

[4] R. Basili, A. Moschitti, and M. T. Pazienza, "Language sensitive text classification," in *RIAO*, 2000, pp. 331-343.

[5] P. S. Jacobs, "Joining statistics with NLP for text categorization," in *Proceedings of the third conference on Applied natural language processing*, 1992, pp. 178-185.

[6] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine learning,* vol. 39, pp. 135-168, 2000.

[7] V. N. Vapnik and V. Vapnik, *Statistical learning theory* vol. 1: Wiley New York, 1998.

[8] G. Salton and M. J. McGill, "Introduction to modern information retrieval," 1986.

[9] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in *Proceedings of the seventh international conference on Information and knowledge management*, 1998, pp. 148-155.

[10] S. M. Weiss, C. Apte, F. J. Damerau, D. E. Johnson, F. J. Oles, T. Goetz*, et al.*, "Maximizing text-mining performance," *IEEE Intelligent systems,* pp. 63-69, 1999.

[11] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM,* vol. 56, pp. 82-89, 2013.

[12] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," in *LREC*, 2010, pp. 2200-2204.

[13] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in *LREC*, 2010, pp. 1320-1326.

[14] T. Blog, "Insights into the #WorldCup conversation on Twitter," in *Twitter Blog*, ed, 2014.

[15] D. Terrana, A. Augello, and G. Pilato, "Automatic Unsupervised Polarity Detection on a Twitter Data Stream," in *Semantic Computing (ICSC), 2014 IEEE International Conference on*, 2014, pp. 128-134.

[16] L. Zhang, "Sentiment analysis on Twitter with stock price and significant keyword correlation," 2013.

[17] A. G. Jivani, "A comparative study of stemming algorithms," *Int. J. Comp. Tech. Appl,* vol. 2, pp. 1930-1938, 2011.

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter,* vol. 11, pp. 10-18, 2009.

[19] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale Bayesian logistic regression for text categorization," *Technometrics,* vol. 49, pp. 291-304, 2007.

[20] H. Daumé III, "Notes on CG and LM-BFGS optimization of logistic regression," vol. 198, p. 282, 2004.

[21] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Machine learning,* vol. 29, pp. 103-130, 1997.

[22] P. Barnaghi, P. Ghaffari, and J. G. Breslin, "Text Analysis and Sentiment Polarity on FIFA World Cup 2014 Tweets," in *Conference ACM SIGKDD 2015*, 2015.