Sentiment Identification in Football-Specific Tweets

Article *in* IEEE Access · December 2018

DOI: 10.1109/ACCESS.2018.2885117

CITATIONS
16

READS
1,519

2 authors:

Samah Aloufi
University of Ottawa
9 PUBLICATIONS **127** CITATIONS

SEE PROFILE

Abdulmotaleb El Saddik
University of Ottawa
722 PUBLICATIONS **9,968** CITATIONS

SEE PROFILE

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 6, 2018, accepted November 26, 2018, date of publication December 5, 2018,
date of current version December 31, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2885117

Sentiment Identification in Football-Specific Tweets

SAMAH ALOUFI AND ABDULMOTALEB EL SADDIK, (Fellow, IEEE)
Multimedia Communication Research Laboratory, School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada
Corresponding author: Samah Aloufi (salou102@uottawa.ca)

ABSTRACT Sports fans generate a large amount of tweets which reflect their opinions and feelings about what is happening during various sporting events. Given the popularity of football events, in this work, we focus on analyzing sentiment expressed by football fans through Twitter. These tweets reflect the changes in the fans' sentiment as they watch the game and react to the events of the game, e.g., goal scoring, penalties, and so on. Collecting and examining the sentiment conveyed through these tweets will help to draw a complete picture which expresses fan interaction during a specific football event. The objective of this work is to propose a domain-specific approach for understanding sentiments expressed in football fans' conversations. To achieve our goal, we start by developing a football-specific sentiment dataset which we label manually. We then utilize our dataset to automatically create a football-specific sentiment lexicon. Finally, we develop a sentiment classifier which is capable of recognizing sentiments expressed in football conversation. We conduct extensive experiments on our dataset to compare the performance of different learning algorithms in identifying the sentiment expressed in football related tweets. Our results show that our approach is effective in recognizing the fans' sentiment during football events.

INDEX TERMS Sentiment analysis, football, soccer, domain-specific, dataset, sentiment lexicon, machine learning, data mining, social media.

I. INTRODUCTION

Sentiment analysis is a growing field of study which has recently seen a spike in its popularity among researchers. Sentiment analysis is the task of identifying user opinions or sentiment regarding an entity [1]. The dramatic growth of sentiment analysis coincides with the increased prominence of social media applications, e.g. Twitter, all of which allow people to actively share their opinions and thoughts towards a vast variety of topics. Activities include, but are not limited to, leaving reviews or opinions about products, events, movies, politics, or other services. This leads to an accumulation of tremendous amounts of opinionated data on social media. Mining this valuable information will benefit a wide range of applications [2]. For example, companies could track consumer opinions towards their products in order to garner information about customer satisfaction levels and to identify which aspects of the products should be improved. Also, this information could be used to compare consumer sentiment about competitive products or services providers. Furthermore, the utility of social media data could be expanded as a representative tool of real-time experiences such as sporting events. Over the course of these events, people generate a massive amount of posts expressing their opinions about the circumstances which occur during sport games. Analyzing the sentiment conveyed in users' posts during football games can be beneficial in indicating if a game has caused a very negative emotion among fans, and could be used to warn the authorities of possible riots after the match [3]. Accordingly, in this work, we focus on the sentiment analysis in football related conversations on Twitter.

Football (soccer) events such as the FIFA World Cup and the UEFA Champions League attract millions around the world. Football, the most popular sport among 8000 different types, boasts 3.5 billion fans all over the world,[1] and has approximately 265 million players.[2] Football's success in garnering the attention of the public and the media is notable. During the FIFA World Cup 2018, fans posted 900 million tweets related to the event.[3] Moreover, Twitter official blog stated that during the FIFA World Cup 2018, 115 billion tweets where people react to what occurred during the

[1] http://www.topendsports.com/world/lists/popular-sport/fans.htm
[2] https://www.fifa.com/mm/document/fifafacts/bcoffsurv/
[3] https://footballcitymediacenter.com/news/20180613/2403377.html

competition such as goals scoring, players injuries or predicting the outcome of the next matches were recorded.[4] This proves that Twitter has become a venue for football fans to discuss and express opinions and emotions regarding the strengths, weaknesses, and events which occur during matches with respect to their team and its opponent. The sentiment analysis and summarization of such large-scale emotional reactions will give us an opportunity to get an insight into how people react to emotionally intensified events such as the win or loss, and how the sentiments change as the events unfold over time. The challenge with the analysis of this large amount of data arises from the fact that unlike traditional documents which are structured and well-written, social data are written in informal languages which include slang and abbreviations, are also affected by spelling and grammar errors, and, they are short in length [4]. In addition, football fans have sport talks on social media. These football talks may sound like a new language for non-fans, especially if combined with slang terms. Furthermore, sentiment expressed by football fans is often accompanied by the use of expletives which make the analysis even more challenging [5]. From a sentiment analysis perspective, using a standard sentiment classifier with football conversations could lead to learning confusion. For instance, ''That long bomb was sick!'' indicates a positive sentiment in the football domain even though the words ''bomb'' and ''sick'' are associated with negative sentiment in general context. Similarly, this tweet: ''Barcelona did it!!!! Holy s***!!!! VIVA BARCELONA!!!! #ChampionsLeague'' in a general sentiment model would be classified as negative because it contains expletives yet it conveys a positive sentiment where the fan is cheering for his team. Thus, the question remains, how do we efficiently analyze sentiment expressed by football fans and track the changes during events time?

Prior studies have been attempted by following machine learning and lexicon-based approaches which are typically used in sentiment analysis such as [6], [7], and [8] which are thoroughly explained in the related work. Fundamental issues appearing in most of the existing works include training sentiment classifiers on general dataset and extracting features based on general sentiment lexicons. The quality of the classification performance is dependent on determining a good set of features. This is especially true for lexical features, since one word or sentence may reflect different sentiments within different domains [9]. Additionally, the lack of a sufficient manually labeled football dataset resulted in a limited progress in football-specific sentiment analysis on social media. Reviewing the literature shows that the only publicly available football sentiment dataset is the FIFA World Cup 2014 introduced by [6] and [7]; however, this dataset is automatically labeled. In this case, the data are prone to a noisy label, where a tweet could be mislabeled and assigned to an incorrect class. For example, this tweet from the FIFA World Cup 2014 dataset ''@FraseForster your a classssss goal keeper'' is labeled as a negative tweet when it actually carried positive sentiment.

In this paper, we intend to address the problem of sentiment analysis of football-specific social conversations. We propose a football-specific sentiment dataset which consists of tweets related to popular football events, and is annotated manually by human beings. We collect data from Twitter which is related to the FIFA World Cup 2014 and the UEFA Champions League 2016/2017 events. Each tweet in our dataset is manually labeled by sentiment (positive, negative, or neutral). To the best of our knowledge, this is the first work on the construction of a manually labeled football sentiment dataset.

We summarize the major contributions of our work as follows:

- We propose a benchmark dataset designed for football tweets sentiment analysis. Our dataset consists of tweets collected from two popular football events: the FIFA World Cup 2014 and the UEFA Champions League 2016/2017. Each tweet in our dataset is manually labeled by sentiment (positive, neutral, or negative) by four annotators. Our dataset consists of 54,526 labeled tweets.
- We develop a football-specific sentiment lexicon that is automatically generated using a corpus-based approach. The lexicon is created using our manually labeled dataset and consists of 3,479 words.
- We conduct a comprehensive experiment to evaluate the performance of different machine learning algorithms and features in identifying sentiment expressed in football related tweets using our proposed football-specific sentiment dataset.

The rest of the paper is organized as follows: related works are discussed in Section II. In Section III, we provide the description of our data collection and annotation. In Section IV, we describe our sentiment analysis approach. Section V provides details on experiments and discusses the results. Finally, Section VI summarizes our findings and possible future work.

II. RELATED WORK

Few studies have been conducted towards football-specific sentiment analysis. Barnaghi *et al.* [6] utilized a logistic regression algorithm to learn polarity (positive and negative) classifier based on Uni-gram and Bi-gram features. They achieved an accuracy of 72% using the Uni-gram feature. In a similar manner, Barnaghi *et al.* [7] combined N-gram features with external lexicon-based features to improve the performance of the Bayesian logistic regression classifier. The best performance of their proposed method was obtained through combining Uni-gram and Bi-gram features which resulted in an accuracy of 74%. Both [6] and [7] first used manually labeled tweets (4,162 tweets) to build their sentiment model. Then they collected tweets during the FIFA World Cup 2014 where they utilized the learned sentiment classifier to find correlation between sentiment and major

[4]https://blog.twitter.com/official/en_us/topics/events/2018/2018-World-Cup-Insights.html

events occurred during the competition. Alves *et al.* [10], proposed sentiment analysis method for football related tweets written in Portuguese. In order to train their sentiment models, they collected tweets during the 2013 FIFA Confederations Cup. The tweets are labeled based on two methods: automatically based on the positive and negative emoticons included in the tweets, and a random sample of 1,500 tweets which were manually labeled. The best performance was achieved by SVM (accuracy of 87%) when trained and tested on the automatically labeled data. Yet, this accuracy dropped to 66% when trained and tested on the manually labeled dataset. On other hand, Gratch *et al.* [8] used lexicon-based features to train Na$^{\prime}$ive Bayes algorithm on SemEval 2014 dataset [11]. They considered the problem of sentiment analysis as classifying a tweet into positive, negative or neutral classes. They proposed training the sentiment classifier on a manually-labeled general dataset, then using the model to identify sentiment in football related tweets. To evaluate the sentiment model performance on football data, they manually labeled 154 tweets related to the FIFA World Cup 2014. Their results showed a strong correlation between the ground truth and the algorithm results. However, the validation set consists of such a small number of tweets, it may not reflect the overall performance of the sentiment model on football related tweets. Aloufi *et al.* [12] trained the SVM classifier on the FIFA World Cup 2014 dataset utilizing N-gram features and different lexicon-based features. The dataset used for training in [12] is automatically labeled, which means it is vulnerable to incorrect labeling. Their proposed method achieved an accuracy of 85% in classifying tweets into positive, negative, or neutral classes.

Previous studies in football sentiment analysis follow the machine learning approach or lexicon-based approach, which are widely used in sentiment analysis. The utilized sentiment lexicons in previous works such as [8] and [7] are general sentiment lexicons which are generated from different resources and domains. Sentiment analysis is dependent on the domain in which it is applied because a word could convey different sentiment and meaning in various domains [13], [14]. This shows the need to develop a football-specific sentiment lexicon. We are only aware of the one football sentiment lexicon, which is constructed based on the automatically labeled FIFA Wolrd Cup dataset, and is introduced in [12]. A machine learning approach relies on a labeled dataset in order to train a classification model. Publicly available sentiment datasets can be divided into: general datasets such as [15], [16], and [17] and domain-specific datasets. The domain-specific datasets include the Health Care Reform dataset [18], the Sanders[5] dataset which consists of tweets from four different topics: Apple, Microsoft, Google and Twitter, the Dialogue Earth Twitter Corpus[6] which consists of three datasets: WA and WB for weather, and GASP for gas price. For the football domain, the publicly available dataset designed for football tweets sentiment analysis is the FIFA World Cup 2014. The FIFA World Cup 2014 dataset, introduced by [6] and [7], consists of 30 million tweets collected during the game time. Each tweet was automatically labeled as either positive, negative or neutral using Aylien Text analysis API.[7] Although the FIFA 2014 dataset is the first sentiment dataset designed specifically for football events, it is automatically labeled utilizing general sentiment algorithm. It is our position that we cannot rely on automatically labeled dataset for accurate results due to the noisy labeling produced by this approach of annotation.

In order to overcome the limitation in previous works, we propose to develop a football-specific sentiment classifier that can effectively recognize sentiment in football related tweets. To do that, we propose a new football dataset which is manually labeled to support the research in sentiment analysis for football tweets. In addition, we develop a new sentiment lexicon using a corpus-based approach. This idea is first introduced in our previous work [12], which considered building a sentiment lexicon and classifier utilizing an automatically labeled football dataset. In this paper, we take a step further by creating a manually labeled dataset specifically for the football domain and comparing the performance of different learning algorithms utilizing a variety of features.

III. FOOTBALL-SPECIFIC SENTIMENT DATASET

In this section, we provide a detailed description of our data collection process and data annotation procedure. The goal of this dataset is to provide a benchmark dataset for the football domain where researchers can utilize this dataset for sentiment analysis and comparison purposes.

A. DATA COLLECTION

We use Twitter as our data source in building our corpus. To ensure that tweets are related to football games, we have collected tweets that were posted during two popular football events: the FIFA World Cup 2014 and the UEFA Champions League 2016/2017.

1) FIFA WORLD CUP (FIFA) 2014

The FIFA World Cup 2014 dataset was collected from Twitter by [6] and [7] during the period between June 6th and July 14th 2014, using the Twitter Streaming API. The tweets were filtered by official hashtags, teams hashtags and user names of teams and players. Each tweet was automatically labeled by polarity using Aylien API.[8] The resulting list of tweets only included the tweets ids and the polarity labels. Thus, we used the Twitter Search API to retrieve the tweets information. Accordingly, we obtained 440,917 English tweets.

[5]http://www.sananalytics.com/lab
[6]www.dialogueearth.org
[7]https://aylien.com/text-api/
[8]https://aylien.com/text-api/

2) CHAMPIONS LEAGUE (CL) 2016/2017

To collect tweets related to CL 2016/2017, we identified the official hashtag '#Championsleague' as a seed to retrieve tweets related to this event. The Twitter Search API was used for the period of June 1st 2016 to June 15th 2017, which is the duration of the targeted event. As a result, a total of 380,579 tweets in multi-languages were obtained. To enrich our data coverage of CL event, we ranked the hashtags that appear in our dataset based on their occurrence in tweets. Thereafter, the top 15 ranked hashtags were selected to retrieve more data from Twitter in the same period of time. Consequently, we collected 2,811,833 tweets, where 37% of tweets are in English, 25% of tweets are in Spanish and 12% of tweets are in French. In our work, we only considered English tweets for sentiment analysis. To ensure data quality, we filtered out duplication and discard retweets which are indicated by the presence of RT symbol. This is left us with 819,848 English tweets.

B. ANNOTATION PROCESS

We annotated our football tweets using crowdsourcing. This platform for creating a benchmark dataset for different tasks is fast, cheap and scalable [19]. Moreover, the accuracy is close to the agreement among experts as indicated in the previous study [20]. We used the Figure Eight service,[9] previously known as CrowdFlower, to crowdsource the annotation of the FIFA 2014 and the CL 2016/2017 tweets. We randomly sampled 25,000 and 31,000 tweets from the CL and the FIFA datasets respectively. In the following subsections, we describe the annotation task design, quality control, and annotation results.

1) TASK DESIGN

The task consisted of reading a tweet related to the FIFA World Cup 2014 or the UEFA Champions League 2016/2017, and evaluating the author sentiment expressed in the tweet (positive, negative, or neutral). The annotators (contributors) were provided with the image URL if it was included in a tweet since it can provide more description. Also, we asked the contributors to select their confidence level in their answer on a 5-point scale (not confident to very confident). For reference, we provided the contributors with examples that show tweets with negative, positive, and neutral sentiment. We also required that all the contributors be familiar with football in order to understand the sentiment that appears in each tweet. We posted 56 jobs on Figure Eight and each job consisted of 1,000 tweets to be annotated by four contributors. Each job contained tweets from either the FIFA or CL, to ensure consistency.

2) QUALITY CONTROL

Quality control plays a fundamental role in fine tuning the annotation process and providing high-quality annotated data [21]. To ensure the quality of the annotation process, we interspersed manually labeled test tweets, known as gold questions on Figure Eight, within other tweets during the annotation process, and monitored the contributors' performance. Each contributor was given an accuracy score that reflected their accuracy on test questions. When a contributor answered a test question incorrectly, he/she was notified and provided with the right answer immediately. Each contributor was expected to maintain an 80% accuracy score during the whole annotation process. If an individual's accuracy fell below the 80%, the contributor was identified as unreliable and was eliminated from the annotation process. Also, all the answers provided by untrusted annotators were discarded. To reduce annotation bias, we only allowed each contributor to annotate a maximum of 10% of the tweets per job. In this manner, the annotation process was not dominated by a small group of contributors. To ensure that annotators have experiences in annotation tasks, we only enlisted users who had achieved level 2 in experience based on Figure Eight standard.

3) ANNOTATION AGGREGATION

We measured the inter-annotation agreement using the average of the pairwise agreement between the annotators. We obtained an agreement of 51% for the CL 2016/2017 dataset and 55% for the FIFA dataset. In constructing the final dataset, we assigned tweets to sentiment categories based on the annotators' agreement on the category and, subsequently, discarded noisy tweets. The football-specific dataset consists of 54,526 tweets in total, where 24,461 tweets belong to CL 2016/2017, and 30,065 tweets are from FIFA 2014. The distribution of tweets among the three sentiment classes are illustrated in Table 1. Our dataset is publicly available for researchers through the MCRLab website.[10]

TABLE 1. Statistics of manually annotated Football-Specific sentiment dataset.

Dataset	#Positive	#Negative	#Neutral	Total
FIFA 2014	13,871	9,648	6,546	30,065
CL 2016/2017	9,562	8,663	6,236	24,461
CL and FIFA	23,433	18,311	12,782	54,526

IV. SENTIMENT ANALYSIS METHOD

The purpose of sentiment analysis is to identify the sentiment of a given text. In general, sentiment analysis can be divided into three levels: document level, sentience level, and fine-grained level. In our work on sentiment analysis, we focus on sentience level, where the goal is to determine whether a tweet conveys a positive, negative or neutral sentiment or opinion [22]. Sentiment analysis can be considered a document classification problem, aimed at separating documents which express positive and negative sentiments

[9] https://www.figure-eight.com/

[10] http://www.mcrlab.net/datasets/

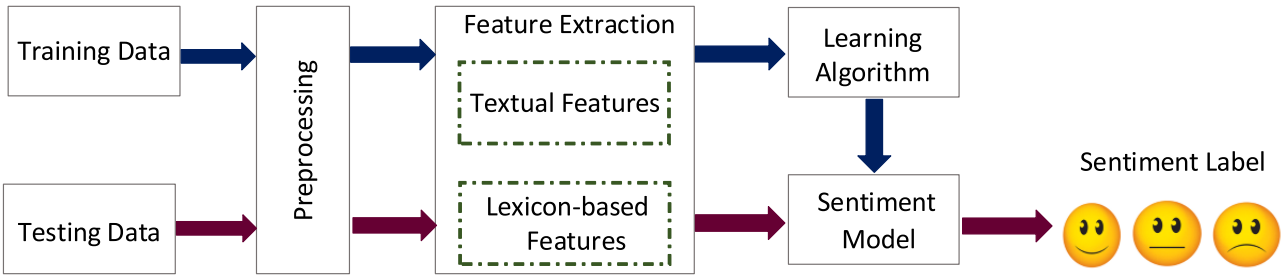


FIGURE 1. General framework for sentiment analysis.

by exploiting certain syntactic and linguistic features [23]. In the literature, the sentiment analysis follows the general framework for the classification problem, which is depicted in Figure 1. Feature extraction and classifier learning are the main components. Feature extraction is the primary step which converts the raw textual data into representative features vectors. These features are then fed into a learning algorithm to learn the classification model. In the following sections, we provide the details of the features adopted in our work and briefly introduce the learning algorithms that are used to learn the sentiment classifier.

A. FEATURES

In our work, we utilize different types of features, such as the Bag-of-Words feature, which is typically used in sentiment analysis. We extract lexicon-based features using various existing general sentiment lexicons. We also develop a new sentiment lexicon oriented for football-specific data, and extract features based on it. The following subsections present the features utilized in our work and the process of creating our football-specific sentiment lexicon.

1) BAG-OF-WORDS (BOW)

BOW is one of the most popular representations of textual data and is widely used in text classification. Given a predefined set of vocabulary $V = \{w_1, w_2, \ldots, w_n\}$, generated using a word or a sequence of words, a document $d \in D$ is represented as an N-dimensional feature vector $X = \{x_1, x_2, \ldots, x_n\}$. Each element x_i in the feature vector corresponds to a word w_i in the vocabulary. The value of x_i can be a binary value that indicates the appearance of a word in the document d_i or the number of occurrence of w_i in d_i that indicates its term frequency (TF). Term Frequency-Inverse Document Frequency (TF-IDF) is another feature representation that reduces the weight assigned to more frequent words appear in the documents' collection. TF-IDF is a popular and successful representation that shows improvement over TF and is calculated as shown in equation 1:

$$TF - IDF_{(w_i)} = TF_{(w_i)} \times log \frac{|D|}{DF_{(w_i)}} \quad (1)$$

where $TF_{(w_i)}$ is the number of occurrence of w_i, $|D|$ is the number of documents in the corpus, and $DF(w_i)$ is the number of document containing the term w_i.

Despite the simplicity and efficiency of BOW, it ignores the co-occurrence of words in texts. To incorporate the word-order, the BOW representation is extended to N-gram language model. In N-gram model, the document is represented as N consecutive words extracted from the collection of documents [24]. The common setting of N-grams is $n \leq 3$. In our work, we investigate the impact of using a 2-gram (Bi-gram), 3-gram (Tri-gram), and a combination of different grams: Uni-gram+Bi-gram, Uni-gram+Tri-gram, and Bi-gram+Tri-gram. We use TF-IDF schema for features vector representation.

2) PART-OF-SPEECH (POS)

POS features are commonly used in sentiment analysis. POS taggers determine the part of speech of each word in a sentence and labels it as noun, verb, adjective…etc. Each tweet is tokenized and part-of-speech tagged using the GATE[11] Twitter POS tagger tool [25]. The frequency of each POS tag is used as a feature vector.

3) EXISTING SENTIMENT LEXICONS

Various sentiment lexicons have been developed using diverse resources; however, the limited coverage of sentiment words in a single lexicon is one of the major limitations. Moreover, large numbers of existing lexicons do not contain the abbreviations, emoticons, and slang widely used in social media. Thus, to achieve better coverage of sentimental words we use features based on the following sentiment lexicons:

- Bing Liu's Opinion Lexicon (OP) [26]: Manually constructed lexicon from customer reviews about various types of products. It has a list of 2,006 positive and 4,781 negative words.
- AFINN-111 Lexicon (AFINN) [27]: Based on Affective Norms for English Words (ANEW). It provides sentiment score for 2,477 words. The score range from 1 to 5 for positive word and from −5 to −1 for negative ones.
- NRC Hashtag Sentiment Lexicon (NRC) [28], [29]: Automatically generated lexicon from 775,000 tweets. The tweets are automatically labeled based on the existing of positive or negative hashtag.
- The Multi-Perspective-Question-Answering Lexicon (MPQA) [30]: It consists of 8,222 words collected from

[11] https://gate.ac.uk/

several resources. Each word is manually labeled with its polarity and intensity.

- Emoticons and Slang Lexicon (Emoticons) [31]: It includes emoticons, social media slang, and abbreviation that are used to express emotion on social media. The total number of entries in this lexicon is 404 and each term is manually labeled as positive or negative.

The lexicons mentioned above are used to extract two features from each individual lexicon: 1) the number of positive tokens and 2) the number of negative tokens.

B. NEW FOOTBALL-ORIENTED SENTIMENT LEXICON

Sentiment lexicons are either constructed manually or automatically. Manually generated lexicons leverage the knowledge of experts to label words or terms based on its sentiment polarity. This method of constructing sentiment lexicon is costly and has limited coverage [32]. Automatic approaches for generating sentiment lexicon can be divided into: corpus-based or thesaurus-based. The thesaurus-based method relies on the assumption that synonyms have the same polarity [33], [34]. The idea of the thesaurus-based method is to begin with seed words where the sentiment orientation is known and then expand the list by including synonyms and antonyms of the seed words [14]. The corpus-based method uses a domain-specific dataset, instead of the dictionary, to create a sentiment lexicon.

To build our football-specific lexicon, we follow the corpus-based approach, using our collection of football-related tweets. First, we preprocess the tweets, removing stop words, URLS, mentions, hashtags, punctuation marks, and digits. In addition to that, we remove key words that are highly correlated with football events such as football clubs' names and their abbreviations, club nick names, and event official hashtags such as "#WorldCup." In generating our lexicon, we consider tweets that belong to positive or negative classes, while neutral tweets are excluded. We then create a vocabulary set that consists of unique tokens appearing in our dataset. The associated sentiment score for each word in the vocabulary set is calculated by adapting the information theoretic approach proposed in [14] and [35]. It is based on the well-known information theoretic measure (TF-IDF) which evaluates the importance of a word in a textual content. The overall score of a word w_i is calculated using equation (2):

$$Score_{(w_i)} = (pos(w_i) - neg(w_i)) \times IDF(w_i)$$

where

$$IDF(w_i) = log\frac{N}{df(w_i)} \tag{2}$$

The overall score of a word w_i is the difference between its positive and negative score multiplied by its inverse document frequency $IDF(w_i)$. N is the total number of tweets in both positive and negative classes, and df is the document frequency (the number of tweet in which w_i appears). Since we have unbalanced classes, we compute $pos(w_i)$ and $neg(w_i)$ based on its frequency relative to positive or negative class as shown in equations 3 and 4:

$$pos_{(w_i)} = \frac{freq(w_i, pos)}{N_(pos)} \times N \tag{3}$$

$$neg_{(w_i)} = \frac{freq(w_i, neg)}{N_(neg)} \times N \tag{4}$$

Our final lexicon, to which we refer as the Football-specific sentiment lexicon, has entries of 3,479 words: 1,422 of which are labeled as positive and 2,057 negative.

As with the existing sentiment lexicon, we extract features which count the number of positive and negative words in each tweet from the football-specific lexicon.

C. LEARNING ALGORITHMS

In this section, we introduce different learning algorithms, including the Support Vector Machine, Naïve Bayes, and Random Forest, all of which are widely used in text classification.

1) SUPPORT VECTOR MACHINE CLASSIFIER (SVM)

SVM algorithm has shown a robust performance in a wide variety of applications, including text classification [36]. The goal of SVM is to select a hyperplane that maximizes the margin between closest instances of the two classes. To find the optimal hyperplane, the following optimization problem needs to be solved:

$$\begin{aligned} &arg\min_{w,b} \frac{1}{2}\|w\|^2 \\ &\text{subject to } y_i(w \cdot x_i + b) \geq 1, \quad i \in [1, n] \end{aligned} \tag{5}$$

where x_i is the feature vector and $y_i \in \{+1, -1\}$ is the label of instance i. In our experiments, we adopt a linear kernel SVM.

2) MULTINOMIAL NAÏVE BAYES CLASSIFIER (MNB)

Naïve Bayes is a probabilistic classifier that despite its simplicity, performs well in text categorization [37]. Given a set of classes $C = \{c_1, c_2, \ldots, c_n\}$ and a set of document $D = \{d_1, d_2, d_3, \ldots, d_m\}$, and $F = \{f_1, f_2, f_3, \ldots, f_k\}$ is the set of features that represent a document $d \in D$. The probability of document d_i belongs to a class c is computed using Bayes' rules presented in equation 6:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \tag{6}$$

The probability of document d is belonging to each class $c \in C$ is calculated individually, then the document d is assigned to class c with the maximum a posteriori (MAP) class as shown in equation 7.

$$c_{(map)} = \arg\max_{c \in C} P(c|d) = \arg\max_{c \in C} \frac{P(d|c)P(c)}{P(d)} \tag{7}$$

Naïve Bayes classifier is referred to as naïve because it assumes that each feature f_i in a document d is conditionally independent from other features in the given document. The denominator $P(d)$ is constant given the input in equation 7

and does not change so we can drop $P(d)$. Thus, we can write equation 7 as the following [38], [39]:

$$c_{(map)} = \arg\max_{c \in C} P(c|d) = \arg\max_{c \in C} p(d|c)P(c)) \quad (8)$$

$$c_{(map)} = \arg\max_{c \in C} p(d|c)P(c)) = \arg\max_{c \in C} P(f_1, f_2, \ldots, f_k|c)P(c) \quad (9)$$

The final equation for Naïve Bayes classifier is defined as in 10:

$$c_{(map)} = \arg\max_{c \in C} P(c) \prod_{f \in F} P(f|c) \quad (10)$$

The Multinomial Naïve Bayes (MNB) classifier is a variant of Naïve Bayes classifier which is used with discrete features such as the counts of words in text classification problem. In our work we adopt the Multinomial Naive Bayes classifier for the sentiment analysis problem.

3) RANDOM FOREST (RF)

Random Forest (RF) is an ensemble classifier that consists of a collection of fully grown decision trees. Each tree in RF is built based on bootstrapped training samples and a random number of features. The overall prediction of the RF is based on the majority votes from all the individual trees [40], [41]. RF shows robust performance to noise and overcomes the over-fitting problem that affects a single decision tree [42].

V. EXPERIMENTS AND RESULTS

Our objective in this work is to build a sentiment classifier that can identify sentiment expressed, specifically, in football tweets. We have utilized our proposed dataset to train different learning algorithms and compare their performance in different experimental settings. In this section, we present the experiment settings and the results of utilizing different learning algorithms.

A. EXPERIMENTAL SETTINGS AND EVALUATION CRITERIA

We conduct experiments in two scenarios: binary and multi-class classification. In the binary classification setting, we ignore the neutral class and consider only tweets with positive and negative polarity. For both settings, we use the CL 2016/2017, FIFA 2014, and FIFA-CL set where we combine tweets collected during both the Champions league and World Cup events. Each dataset is randomly divided into 60%-40% for training and testing, respectively. Accuracy and F-score metrics are used to evaluate the performance of different learning algorithms. Accuracy is defined as shown in equation 11 and F-score is calculated as illustrated in equation 12.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (11)$$

where TP, TN, FP, and FN refer to true positive, true negative, false positive and false negative.

$$F\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (12)$$

where precision is calculated as $\frac{TP}{TP+FP}$ and recall is defined as $\frac{TP}{TP+FN}$.

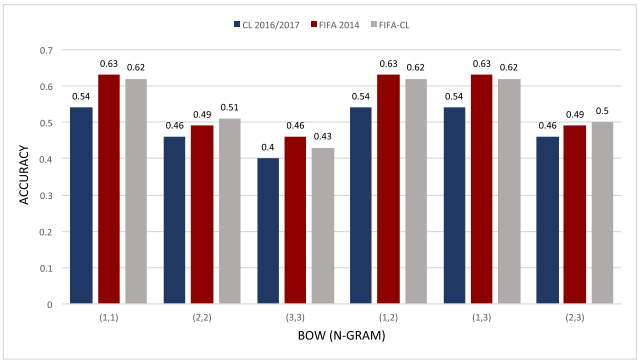
B. EXPERIMENTAL RESULTS

In this section, we present empirical results that demonstrate the effect of using different features on the detection of a given tweet sentiment. In section V-B.1, we report the results of using three classes, and in section V-B.2 we discuss the results of using the binary classification setting. Furthermore, in section V-B.3 we investigate the generalization capability of the sentiment model using cross-dataset setting.

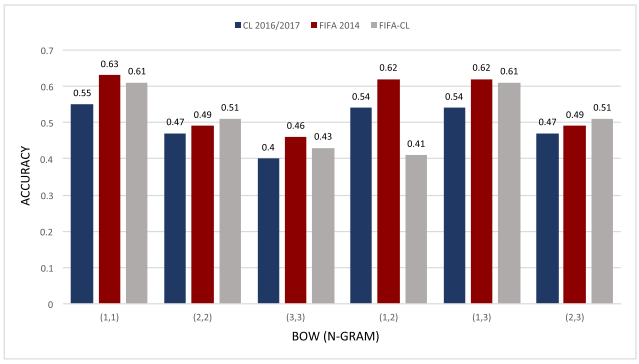
1) MULTI-CLASS CLASSIFICATION RESULTS

In this section, we will discuss the performance of different features in assigning a tweet to one of the three sentiment classes: positive, neutral, or negative. First, we investigate the impact of BOW and N-gram models on the performance of SVM, MNB, and RF sentiment classifiers. The results are illustrated in Figures 2 and 3 in accuracy and F-score, respectively. The experimental results show that Uni-gram has achieved better performance than Bi-gram and Tri-gram when used individually. The SVM algorithm has obtained an accuracy of 63% when trained on the FIFA 2014 dataset using Uni-gram feature. This accuracy decreased to 49% and 46% using Bi-gram and Tri-gram respectively. From the results, we can see that the Tri-gram model performance is the worst with respect to other N-gram models. This is due to the limited number of a tweet's characters and the fact that frequent appearances of a Tri-gram are rare in tweets. The combination of Bi-gram+Uni-gram, and Tri-gram+Uni-gram has boosted the performance of Bi-gram and Tri-gram models with respect to accuracy and F-score of SVM, MNB and RF classification models, and among the three datasets. In comparing the performance of SVM, MNB and RF, the results show that SVM has achieved the best performance. The difference in performance accuracy and F-score between SVM and MNB is not significant. This observation is consistent among CL 2016/2017, FIFA 2014, and FIFA-CL datasets. The N-gram models, in general, perform better on FIFA 2014 and FIFA-CL datasets than on the CL 2016/2017, where they include more training instances. Therefore, having sufficient data for training is important for improving the classification model performance.

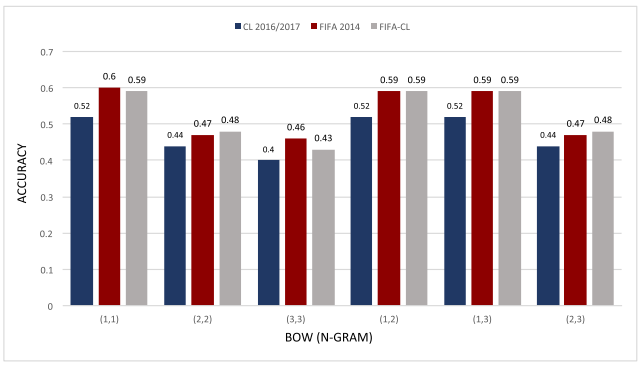
Table 2 illustrates the results of different lexicons and POS features. We also include the best results from the BOW model for comparison purposes. Among all the features, the AFINN lexicon has achieved the worst performance in terms of accuracy and F-score. Similarly, the Emoticons and NRC lexicons show lower performance levels than Opinion and Football lexicons, since many tweets do not include emoticons or explicit emotional hashtags. The Opinion lexicon outperformed other general sentiment lexicons. When comparing the performance of the Football lexicon to other general lexicons, the results illustrate that the Football lexicon achieves similar performance when compared to the



(a)

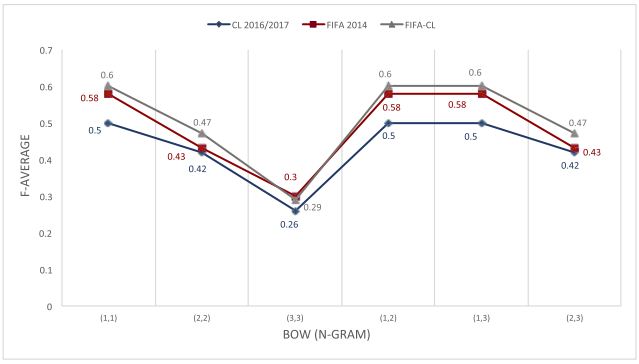


(b)

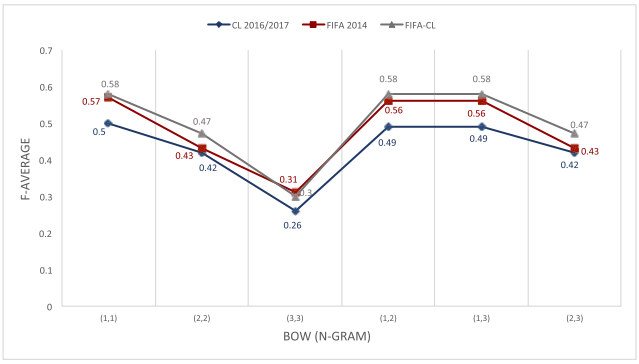


(c)

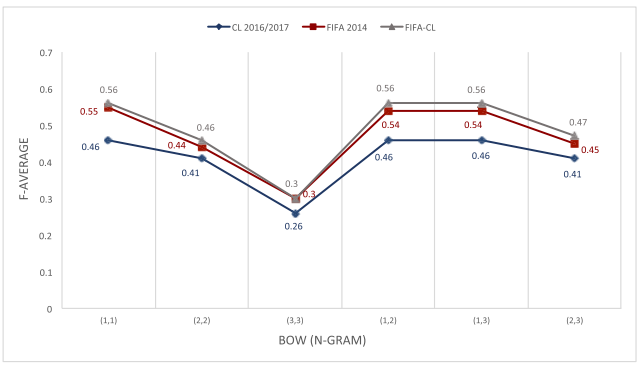
FIGURE 2. Accuracy of different BOW (N-gram) models on CL 2016/2017, FIFA 2014, and FIFA-CL datasets for multi-class classification setting. In (a) the results of SVM classifier, (b) the results of MNB classifier, and in (c) the results of RF classifier.



(a)



(b)



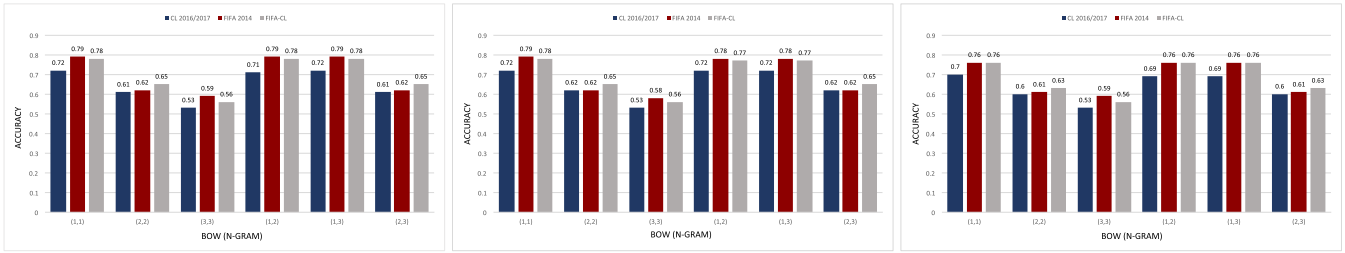
(c)

FIGURE 3. Average F-score of different BOW (N-gram) models on CL 2016/2017, FIFA 2014, and FIFA-CL datasets for multi-class classification setting. In (a) the results of SVM classifier, (b) the results of MNB classifier, and in (c) the results of RF classifier.

best general sentiment lexicon (Opinion lexicon) used in our experiment. Using the Football lexicon, the MNB classifier has obtained an accuracy of 56% whereas the accuracy has dropped to 51% using the Opinion lexicon on the FIFA 2014 dataset. For POS feature, it outperforms the AFINN lexicon in terms of accuracy and F-score. In some cases, the POS feature shows better performance than the NRC and the Emoticons lexicons. The best results when investigating the performance of different features on the three datasets individually comes from the BOW (Uni-gram model). Comparing the performance of the learning algorithms on the CL 2016/2017, the FIFA 2014, and the FIFA-CL datasets shows that the SVM outperforms the MNB and RF classifiers.

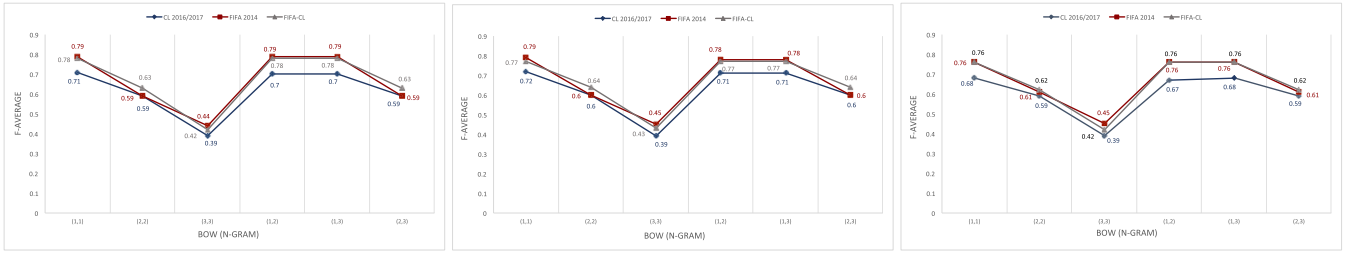
Through the process, we observed that the performance of the different learning algorithms are better on larger datasets. For example, the SVM classifier has achieved an accuracy of 54% on the CL 2016/2017 dataset using BOW. This accuracy increases to 63% on the FIFA dataset.

We have presented the sentiment performance results through the exploration of individual features. Now, we will provide the results of combining different features in order to boost the algorithms' performance. We examine the impact of fusing BOW, lexicons and POS features. The results are illustrated in Table 3. Combining features extracted from different general sentiment lexicons boosts the SVM algorithm



(a) (b) (c)

FIGURE 4. Accuracy of different BOW (N-gram) models on CL 2016/2017, FIFA 2014, and FIFA-CL datasets for binary classification setting. In (a) the results of SVM classifier, (b) the results of MNB classifier, and in (c) the results of RF classifier.



(a) (b) (c)

FIGURE 5. Average F-score of different BOW (N-gram) models on CL 2016/2017, FIFA 2014, and FIFA-CL datasets for binary classification setting. In (a) the results of SVM classifier, (b) the results of MNB classifier, and in (c) the results of RF classifier.

TABLE 2. Performance of different features on: CL 2016/2017, FIFA 2014, and FIFA-CL datasets utilizing different learning algorithms.

		CL 2016/2017		FIFA 2014		FIFA-CL	
Classifiers	**Features**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**
SVM	BOW	0.54	0.50	0.63	0.58	0.62	0.60
	Opinion Lexicon	0.52	0.44	0.61	0.53	0.56	0.49
	AFINN Lexicon	0.39	0.22	0.46	0.29	0.43	0.26
	MPQA Lexicon	0.48	0.41	0.56	0.49	0.52	0.45
	NRC Lexicon	0.48	0.41	0.55	0.48	0.51	0.45
	Emoticons Lexicon	0.40	0.24	0.47	0.32	0.44	0.28
	Football Lexicon	0.51	0.43	0.55	0.49	0.55	0.48
	POS	0.43	0.37	0.47	0.42	0.46	0.40
MNB	BOW	0.55	0.50	0.63	0.57	0.61	0.58
	Opinion Lexicon	0.50	0.42	0.51	0.40	0.47	0.35
	AFINN Lexicon	0.39	0.22	0.46	0.29	0.43	0.26
	MPQA Lexicon	0.47	0.39	0.47	0.32	0.44	0.28
	NRC Lexicon	0.40	0.25	0.46	0.29	0.43	0.26
	Emoticons Lexicon	0.40	0.24	0.47	0.32	0.44	0.28
	Football Lexicon	0.51	0.43	0.56	0.48	0.54	0.46
	POS	0.42	0.35	0.46	0.31	0.43	0.28
RF	BOW	0.52	0.46	0.60	0.55	0.59	0.56
	Opinion Lexicon	0.52	0.44	0.60	0.53	0.56	0.49
	AFINN Lexicon	0.39	0.22	0.46	0.29	0.43	0.26
	MPQA Lexicon	0.48	0.41	0.56	0.49	0.52	0.45
	NRC Lexicon	0.48	0.41	0.54	0.48	0.51	0.45
	Emoticons Lexicon	0.40	0.24	0.47	0.32	0.44	0.28
	Football Lexicon	0.51	0.43	0.55	0.49	0.56	0.48
	POS	0.42	0.40	0.44	0.42	0.45	0.44

performance slightly, from obtaining 61% accuracy using only Opinion lexicon to 62%, on the FIFA 2014 dataset. SVM has achieved accuracy of 54% when fusing general and Football lexicons features, compared to 52% using only the Opinion lexicon on the CL 2016/2017 dataset. The combination of BOW and general lexicons features has slightly improved the performance for SVM, MNB, and RF sentiment models.

Similarly, combining BOW and Football lexicon features boosts the performance from 54% accuracy to ≈56% when using the SVM model on the CL 2016/2017 dataset, see Table 3. The best performance, in general, was achieved by combining all the features rather than relying on a single type of feature.

2) BINARY CLASSIFICATION RESULTS

We perform experiments on polarity classification using the positive and negative tweets in our constructed datasets. Figures 4 and 5 report the results of BOW and N-gram models in accuracy and F-score, respectively. The Uni-gram model has showed the best performance among the learning algorithms on the three datasets. The Tri-gram model performance is the worst compared to other N-gram models. Bi-gram and Tri-gram models show improvement in their performance when combined with Uni-gram model. For example, the combination of Uni-gram and Bi-gram results in accuracy of 79% of the SVM classifier compared to 62% using only Bi-gram model on the FIFA 2014 dataset. This is similar to the observation in the multi-classification setting where the best performance is obtained by the use of Uni-gram model. From the results, we can see that the learning algorithms perform better in binary classification than in multi-classification task. This is because learning algorithms are affected by sentiment class distribution and perform poorly in rare class.

Comparison of different features' performance to the BOW is shown in Table 4. In comparing lexicon features' performance, Opinion and Football lexicons have achieved the best performance than other lexicons. The MNB algorithm has achieved better results using the Football lexicon than

TABLE 3. Performance of features combination on different classifiers using the three datasets.

	SVM						MNB						RF					
	CL 2016/2017		FIFA 2014		FIFA-CL		CL 2016/2017		FIFA 2014		FIFA-CL		CL 2016/2017		FIFA 2014		FIFA-CL	
Features	Accuracy	F-Average	Accuracy	F-Average	Accuracy	F-Average	Accuracy	F-Average	Accuracy	F-Average	Accuracy	F-Average	Accuracy	F-Average	Accuracy	F-Average	Accuracy	F-Average
BOW+ General Lexicon	0.56	0.52	0.64	0.59	0.63	0.61	0.56	0.51	0.64	0.58	0.62	0.59	0.55	0.50	0.63	0.58	0.62	0.59
BOW+POS	0.55	0.51	0.63	0.58	0.63	0.61	0.55	0.50	0.62	0.57	0.61	0.58	0.52	0.46	0.58	0.53	0.59	0.56
POS+General Lexicon	0.52	0.47	0.61	0.55	0.58	0.52	0.51	0.43	0.58	0.51	0.54	0.46	0.52	0.50	0.61	0.57	0.58	0.56
BOW+Football Lexicon	0.55	0.51	0.62	0.57	0.62	0.60	0.55	0.50	0.62	0.56	0.61	0.58	0.53	0.48	0.59	0.54	0.60	0.56
POS+Football Lexicon	0.51	0.44	0.56	0.50	0.56	0.50	0.51	0.44	0.56	0.48	0.55	0.47	0.50	0.47	0.55	0.51	0.55	0.53
General Lexicon	0.52	0.47	0.62	0.54	0.57	0.50	0.50	0.42	0.58	0.51	0.54	0.46	0.51	0.46	0.60	0.53	0.56	0.50
General lexicon+Football lexicon	0.54	0.47	0.61	0.54	0.59	0.51	0.53	0.45	0.61	0.53	0.57	0.50	0.50	0.49	0.58	0.55	0.57	0.55
BOW+POS+General Lexicon	0.57	0.53	0.65	0.60	0.64	0.62	0.56	0.51	0.64	0.58	0.62	0.59	0.54	0.49	0.64	0.58	0.61	0.59
BOW+Football Lexicon+POS	0.56	0.52	0.62	0.58	0.63	0.61	0.55	0.50	0.62	0.56	0.61	0.58	0.53	0.47	0.59	0.53	0.60	0.56
All features	0.57	0.53	0.64	0.59	0.64	0.62	0.56	0.51	0.63	0.57	0.62	0.59	0.55	0.50	0.63	0.57	0.62	0.59

TABLE 4. Performance of different features on: CL 2016/2017, FIFA 2014, and FIFA-CL datasets utilizing different learning algorithms in binary classification setting.

		CL 2016/2017		FIFA 2014		FIFA-CL	
Classifiers	Features	Accuracy	F-Average	Accuracy	F-Average	Accuracy	F-Average
SVM	BOW	0.72	0.71	0.79	0.79	0.78	0.78
	Opinion Lexicon	0.69	0.69	0.77	0.77	0.74	0.73
	AFINN Lexicon	0.52	0.36	0.59	0.44	0.56	0.40
	MPQA Lexicon	0.65	0.64	0.72	0.72	0.68	0.68
	NRC Lexicon	0.65	0.65	0.70	0.70	0.67	0.67
	Emoticons Lexicon	0.54	0.40	0.60	0.47	0.57	0.44
	Football Lexicon	0.68	0.68	0.71	0.72	0.72	0.73
	POS	0.58	0.58	0.60	0.59	0.60	0.59
MNB	BOW	0.72	0.71	0.79	0.79	0.78	0.77
	Opinion Lexicon	0.67	0.65	0.65	0.58	0.62	0.53
	AFINN Lexicon	0.52	0.36	0.59	0.44	0.56	0.40
	MPQA Lexicon	0.63	0.61	0.60	0.48	0.57	0.44
	NRC Lexicon	0.54	0.40	0.59	0.44	0.56	0.40
	Emoticons Lexicon	0.54	0.40	0.60	0.47	0.57	0.44
	Football Lexicon	0.68	0.67	0.71	0.70	0.71	0.69
	POS	0.57	0.55	0.59	0.46	0.57	0.44
RF	BOW	0.70	0.68	0.76	0.76	0.76	0.76
	Opinion Lexicon	0.69	0.69	0.76	0.77	0.73	0.73
	AFINN Lexicon	0.52	0.36	0.59	0.44	0.56	0.40
	MPQA Lexicon	0.64	0.64	0.72	0.72	0.68	0.68
	NRC Lexicon	0.64	0.64	0.70	0.70	0.67	0.67
	Emoticons Lexicon	0.54	0.40	0.60	0.47	0.57	0.44
	Football Lexicon	0.68	0.68	0.70	0.71	0.73	0.72
	POS	0.56	0.56	0.58	0.58	0.59	0.59

the Opinion lexicon. However, the SVM and RF classifiers show better performance using Opinion lexicon. For instance, the SVM obtains accuracy of 74% using the Opinion lexicon compared to 72% when using the Football lexicon on the FIFA-CL dataset. The AFINN lexicon performance is the worst among all other lexicons. It may be a result of limited number of words included in the lexicon. Interestingly, POS feature has provided better performance than AFINN lexicon, and in some cases, it outperformed the Emoticon lexicon. This shows that POS could be helpful in identifying sentiment in a tweet.

Besides exploring the performance of different features individually, we have examined the effect of fusing multiple features on the performance of the learning algorithms. The results are illustrated in Table 5. Compared to the best single general lexicon feature performance, the fusion of the general lexicons features has improved the performance of the sentiment classifiers by 1% in general. In some cases, the accuracy increases by 9% such in the case of MNB algorithm on the FIFA 2014 dataset. The combination of general and Football lexicons features has boosted the performance of SVM, MNB, and RF in identifying sentiment of football tweets. The accuracy of the SVM model increases by 3%, from 69% accuracy achieved using only general lexicon to 72% on the CL 2016/2017 dataset. The MNB classifier shows an increase in the performance accuracy by 7% when combining the general and specific lexicons on the FIFA 2014 dataset and by 4% on the FIFA-CL dataset. The best performance has been achieved in most of the cases, by fusing all the features.

3) CROSS DATASET PERFORMANCE

Cross-dataset sentiment classification is defined as training a sentiment model on a dataset S_i to predict the sentiment of a tweet t_k in a dataset S_j. We conduct a cross-dataset experiment where we have employed the classification models learned from CL 2016/2017 on FIFA 2014 and vice versa. We have used identical experiment settings as in the previous sections: multi-class and binary classifications. The results of multi-classification experiments are illustrated in Tables 6 and 7. Table 6 lists the results of different classifiers trained on the CL 2016/2017 dataset and tested on the FIFA 2014. We can see from the results that the accuracy and F-score of the classifiers: SVM, MNB, and RF learned from the CL 2016/2017, Table 6, surpass the performance of the classifiers learned from the FIFA 2014 dataset. This is because accuracy and F-score are affected by the number of instances in each class. The imbalance between classes is larger in the FIFA 2014 than in the CL 2016/2017 dataset, which impacts the performance of the classifiers in classifying tweets that belong to neutral class.

In comparing different individual features' performances, BOW has showed superior performance in relation to other features. The Football lexicon and Opinion lexicon have exceeded the performance of other lexicons and POS features. Note that the Opinion lexicon has been manually generated, whilst the Football-Specific lexicon has been automatically constructed from football dataset. Likewise, combining the general lexicon features exhibits similar performance in terms of accuracy to the Football lexicon, as the results show in Tables 6 and 7. Nevertheless, combining general and Football lexicons features has boosted the performance of the learning algorithms. The performance of the SVM classifier in Table 6 has demonstrated an accuracy of 61% using general and Football lexicons while it achieved 59% using only general lexicons.

The results of the Binary classification setting is presented in Tables 8 and 9. Comparing the results of each feature individually, we can see that the highest accuracy achieved by the BOW feature. Analyzing the results which achieved by the automatically generated lexicon NRC and Football

TABLE 5. Performance of features combination on different classifiers using the three datasets for binary classification setting.

	SVM						MNB						RF					
	CL 2016/2017		FIFA 2014		FIFA-CL		CL 2016/2017		FIFA 2014		FIFA-CL		CL 2016/2017		FIFA 2014		FIFA-CL	
Features	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**
BOW+General Lexicon	0.73	0.73	0.81	0.81	0.79	0.79	0.74	0.73	0.81	0.81	0.79	0.79	0.73	0.73	0.80	0.80	0.78	0.78
BOW+POS	0.72	0.71	0.79	0.79	0.78	0.78	0.73	0.72	0.79	0.79	0.78	0.77	0.69	0.68	0.74	0.74	0.75	0.74
POS+General Lexicon	0.70	0.70	0.78	0.78	0.75	0.74	0.69	0.68	0.75	0.74	0.70	0.68	0.71	0.71	0.77	0.77	0.75	0.75
BOW+Football Lexicon	0.73	0.72	0.77	0.78	0.78	0.78	0.73	0.72	0.78	0.79	0.78	0.78	0.72	0.71	0.76	0.76	0.77	0.77
POS+Football Lexicon	0.69	0.69	0.72	0.72	0.73	0.73	0.69	0.68	0.71	0.70	0.71	0.71	0.67	0.67	0.71	0.71	0.72	0.72
General Lexicon	0.70	0.69	0.79	0.79	0.75	0.74	0.68	0.66	0.74	0.73	0.70	0.68	0.69	0.69	0.77	0.77	0.74	0.74
General lexicon+Football lexicon	0.72	0.72	0.78	0.78	0.77	0.77	0.71	0.70	0.78	0.78	0.75	0.74	0.70	0.70	0.76	0.76	0.76	0.76
BOW+POS+General Lexicon	0.74	0.73	0.81	0.81	0.79	0.79	0.74	0.74	0.81	0.81	0.79	0.78	0.73	0.73	0.80	0.80	0.78	0.78
BOW+Football Lexicon+POS	0.73	0.73	0.78	0.78	0.79	0.79	0.73	0.72	0.78	0.78	0.78	0.78	0.71	0.71	0.75	0.75	0.76	0.76
All features	0.74	0.74	0.80	0.81	0.80	0.80	0.74	0.74	0.80	0.80	0.79	0.79	0.74	0.74	0.79	0.80	0.79	0.79

TABLE 6. Performance results of different features on classifying sentiment of FIFA 2014 dataset using models learned from CL 2016/2017 dataset.

	SVM		MNB		RF	
Features	Accuracy	F-average	Accuracy	F-average	Accuracy	F-average
BOW	0.61	0.59	0.60	0.57	0.56	0.53
Opinion Lexicon	0.58	0.51	0.57	0.50	0.58	0.51
AFINN Lexicon	0.46	0.29	0.46	0.29	0.46	0.29
MPQA Lexicon	0.54	0.48	0.50	0.40	0.54	0.48
NRC Lexicon	0.54	0.47	0.46	0.30	0.54	0.47
Emoticons Lexicon	0.47	0.32	0.47	0.32	0.47	0.32
Football Lexicon	0.57	0.52	0.58	0.50	0.58	0.51
POS	0.47	0.44	0.48	0.40	0.43	0.43
General Lexicon	0.59	0.53	0.58	0.51	0.58	0.53
BOW+General Lexicon	0.63	0.61	0.62	0.58	0.61	0.61
BOW+POS	0.61	0.60	0.60	0.57	0.56	0.52
POS+General Lexicon	0.60	0.56	0.58	0.51	0.58	0.57
BOW+Football Lexicon	0.61	0.59	0.61	0.57	0.59	0.54
POS+Football Lexicon	0.58	0.53	0.58	0.51	0.54	0.53
General lexicon+Football lexicon	0.61	0.55	0.61	0.53	0.57	0.55
BOW+POS+General Lexicon	0.63	0.62	0.62	0.59	0.62	0.60
BOW+Football Lexicon+POS	0.62	0.61	0.61	0.57	0.60	0.56
All features	0.63	0.63	0.63	0.59	0.63	0.61

TABLE 7. Performance results of different features on classifying sentiment of CL 2016/2017 dataset using models learned from FIFA 2014 dataset.

	SVM		MNB		RF	
Features	**Accuracy**	**F-average**	**Accuracy**	**F-average**	**Accuracy**	**F-average**
BOW	0.57	0.52	0.57	0.52	0.55	0.51
Opinion Lexicon	0.53	0.46	0.41	0.26	0.53	0.46
AFINN Lexicon	0.39	0.22	0.39	0.22	0.39	0.22
MPQA Lexicon	0.49	0.41	0.39	0.22	0.49	0.41
NRC Lexicon	0.48	0.41	0.39	0.22	0.48	0.41
Emoticons Lexicon	0.40	0.24	0.40	0.24	0.40	0.24
Football Lexicon	0.53	0.45	0.49	0.41	0.53	0.46
POS	0.43	0.35	0.39	0.23	0.41	0.39
General Lexicon	0.54	0.46	0.46	0.36	0.54	0.46
BOW+General Lexicon	0.58	0.54	0.58	0.53	0.57	0.53
BOW+POS	0.57	0.54	0.57	0.52	0.53	0.47
POS+General Lexicon	0.54	0.48	0.48	0.39	0.54	0.51
BOW+Football Lexicon	0.57	0.52	0.57	0.52	0.56	0.49
POS+Football Lexicon	0.54	0.47	0.50	0.42	0.52	0.48
General lexicon+Football lexicon	0.56	0.48	0.54	0.46	0.54	0.52
BOW+POS+General Lexicon	0.59	0.55	0.58	0.53	0.57	0.52
BOW+Football Lexicon+POS	0.58	0.54	0.57	0.52	0.56	0.49
All features	0.59	0.55	0.58	0.53	0.58	0.53

TABLE 8. Performance results on FIFA 2014 dataset using models learned from CL 2016/2017 dataset utilizing different features in binary classification setting.

	SVM		MNB		RF	
Features	Accuracy	F-average	Accuracy	F-average	Accuracy	F-average
BOW	0.76	0.76	0.75	0.75	0.72	0.70
Opinion Lexicon	0.75	0.75	0.74	0.73	0.75	0.75
AFINN Lexicon	0.59	0.44	0.59	0.44	0.59	0.44
MPQA Lexicon	0.70	0.70	0.65	0.58	0.70	0.69
NRC Lexicon	0.69	0.69	0.59	0.45	0.69	0.69
Emoticons Lexicon	0.60	0.47	0.60	0.47	0.60	0.47
Football Lexicon	0.74	0.74	0.74	0.73	0.74	0.74
POS	0.61	0.61	0.62	0.58	0.57	0.57
General Lexicon	0.76	0.76	0.75	0.74	0.75	0.75
BOW+General Lexicon	0.79	0.79	0.78	0.78	0.78	0.78
BOW+POS	0.77	0.76	0.76	0.76	0.71	0.68
POS+General Lexicon	0.76	0.76	0.74	0.74	0.75	0.75
BOW+Football Lexicon	0.78	0.78	0.76	0.76	0.76	0.76
POS+Football Lexicon	0.74	0.74	0.74	0.74	0.72	0.72
General lexicon+Football lexicon	0.78	0.78	0.78	0.78	0.76	0.76
BOW+POS+General Lexicon	0.79	0.79	0.78	0.78	0.78	0.78
BOW+Football Lexicon+POS	0.78	0.78	0.77	0.76	0.76	0.76
All features	0.80	0.79	0.78	0.78	0.80	0.79

TABLE 9. Performance results on CL 2016/2017 dataset using models learned from FIFA 2014 dataset utilizing different features in binary classification setting.

	SVM		MNB		RF	
Features	Accuracy	F-average	Accuracy	F-average	Accuracy	F-average
BOW	0.74	0.74	0.75	0.75	0.73	0.73
Opinion Lexicon	0.72	0.72	0.55	0.41	0.72	0.72
AFINN Lexicon	0.52	0.36	0.52	0.36	0.52	0.36
MPQA Lexicon	0.67	0.67	0.53	0.37	0.65	0.64
NRC Lexicon	0.64	0.64	0.52	0.36	0.64	0.64
Emoticons Lexicon	0.54	0.39	0.54	0.39	0.54	0.39
Football Lexicon	0.72	0.72	0.66	0.63	0.71	0.71
POS	0.57	0.54	0.53	0.37	0.56	0.55
General Lexicon	0.73	0.72	0.62	0.57	0.72	0.72
BOW+General Lexicon	0.76	0.76	0.76	0.76	0.76	0.76
BOW+POS	0.75	0.75	0.75	0.75	0.71	0.71
POS+General Lexicon	0.73	0.72	0.65	0.61	0.73	0.72
BOW+Football Lexicon	0.75	0.75	0.76	0.76	0.75	0.75
POS+Football Lexicon	0.72	0.72	0.67	0.66	0.71	0.70
General lexicon+Football lexicon	0.75	0.75	0.72	0.71	0.73	0.73
BOW+POS+General Lexicon	0.77	0.77	0.76	0.76	0.76	0.75
BOW+Football Lexicon+POS	0.76	0.76	0.76	0.76	0.74	0.74
All features	0.77	0.77	0.77	0.77	0.77	0.77

lexicons demonstrates that the specific lexicon (Football) outperforms the general lexicon with respect to accuracy and F-score measures. In addition, the integration of general and Football lexicons features has attained better performance than relying on the combination of general lexicons features or individual lexicon. However, fusing both POS and BOW features with general lexicon has raised the accuracy of the SVM classifier by 1% over the combination with the Football lexicon. Here, it worth to mention that integration of general lexicons includes manually and automatically generated sentiment lexicons and the performance of this combination is very similar to a single automatically constructed specific lexicon. Overall, the best performance of SVM, MNB, and RF classifiers comes from the integration of all the features as illustrated in Tables 8 and 9. As mentioned before, the result of different sentiment models (SVM, MNB, and RF) trained on the CL 2016/2017 show a more robust performance than the ones trained on the FIFA 2014 dataset. Although the FIFA 2014 includes more training tweets, there is a gap in the number of positive and negative tweets that impacts the classifiers' performance. Consequently, balanced data in the number of instances that belong to each class is important in sentiment analysis.

VI. CONCLUSION AND FUTURE WORK

In this work, we have proposed a football-specific sentiment dataset which consists of tweets collected from the FIFA World Cup 2014 and the UEFA Champions League 2016/2017 football events. Our dataset consists of 54,526 tweets manually labeled by four annotators. We have also developed a sentiment lexicon that is oriented for the football domain using corpus-based approach. The Football-Specific sentiment lexicon includes a list of 3,479 words labeled according to its polarity. In our work, we have conducted an extensive experiment to evaluate the performance of three learning algorithms (SVM, MNB, and RF) in recognizing sentiment appearing in football conversations on social media utilizing different features on our proposed dataset. The results have shown that the BOW model (Uni-gram) has achieved the best performance in comparison to other features. Aside from BOW, lexicon-based features have achieved comparable performances to the BOW. Specifically, the Opinion lexicon and the Football-Specific lexicon have obtained better performance levels than other lexicons used in our experiments. The SVM, in general, has demonstrated robust and consistent performance in comparison with MNB and RF. Moreover, the experimental results have illustrated that training classifiers on datasets with a sufficient and balanced number of instances improves the performance of the learning algorithms.

In this work, we consider utilizing the standard features in sentiment and text classification. For future work, we plan to investigate the impact of advanced features such as semantic and contextual features on sentiment model performance; as well as, multimedia sentiment analysis. The sentiment analysis presented in this paper can be extended to generate a subjective summary that describes fans' reactions to events which occur during the games. Besides the sentiment analysis, analyzing emotions that appear in football-related tweets is a possible direction for future research.

REFERENCES

[1] B. Liu, "Sentiment analysis and opinion mining," *Synth. Lectures Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012.
[2] T. Al-Moslmi, N. Omar, S. Abdullah, and M. Albared, "Approaches to cross-domain sentiment analysis: A systematic literature review," *IEEE Access*, vol. 5, pp. 16173–16192, 2017.
[3] D. Stojanovski, G. Strezoski, G. Madjarov, and I. Dimitrovski, "Emotion identification in FIFA world cup tweets using convolutional neural network," in *Proc. 11th Int. Conf. Innov. Inf. Technol. (IIT)*, Nov. 2015, pp. 52–57.
[4] A. Giachanou and F. Crestani, "Like it or not: A survey of Twitter sentiment analysis methods," *ACM Comput. Surv.*, vol. 49, no. 2, p. 28, 2016.
[5] E. Byrne and D. Corney, "Sweet FA: Sentiment, swearing and soccer," in *Proc. ICMR 1st Workshop Social Multimedia Storytelling*, Glasgow, Scotland, 2014.
[6] P. Barnaghi, P. Ghaffari, and J. G. Breslin, "Text analysis and sentiment polarity on FIFA world cup 2014 tweets," in *Proc. Conf. ACM SIGKDD*, vol. 15, 2015, pp. 10–13.
[7] P. Barnaghi, P. Ghaffari, and J. G. Breslin, "Opinion mining and sentiment polarity on twitter and correlation between events and sentiment," in *Proc. IEEE 2nd Int. Conf. Big Data Comput. Service Appl. (BigDataService)*, Mar. 2016, pp. 52–57.
[8] J. Gratch, G. Lucas, N. Malandrakis, E. Szablowski, E. Fessler, and J. Nichols, "GOAALLL!: Using sentiment in the world cup to explore theories of emotion," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2015, pp. 898–903.
[9] N. Malandrakis, M. Falcone, C. Vaz, J. Bisogni, A. Potamianos, and S. Narayanan, "SAIL: Sentiment analysis using semantic similarity and contrast," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, 2014, pp. 512–516.
[10] A. L. F. Alves, A. Luiz, C. de Souza Baptista, A. A. Firmino, M. G. de Oliveira, and A. C. de Paiva, "A comparison of SVM versus naive-Bayes techniques for sentiment analysis in tweets: A case study with the 2013 FIFA confederations cup," in *Proc. 20th Brazilian Symp. Multimedia Web (WebMedia)*, New York, NY, USA, 2014, pp. 123–130.
[11] S. Rosenthal, A. Ritter, P. Nakov, and V. Stoyanov, "Semeval-2014 task 9: Sentiment analysis in Twitter," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, Aug. 2014, pp. 73–80.
[12] S. Aloufi, F. Alzamzami, M. Hoda, and A. E. Saddik, "Soccer fans sentiment through the eye of big data: The UEFA champions league as a case study," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Apr. 2018, pp. 244–250.
[13] Z. Wang, V. J. C. Tong, P. Ruan, and F. Li, "Lexicon knowledge extraction with sentiment polarity computation," in *Proc. IEEE 16th Int. Conf. Data Mining Workshops (ICDMW)*, Dec. 2016, pp. 978–983.
[14] K. Labille, S. Gauch, and S. Alfarhood, "Creating domain-specific sentiment lexicons via text mining," in *Proc. Workshop Issues Sentiment Discovery Opinion Mining (WISDOM)*, Aug. 2017.
[15] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in *Proc. 7th Conf. Int. Lang. Resour. Eval. (LREC)*, 2010, pp. 1320–1326.
[16] A. Go, "Sentiment classification using distant supervision," Stanford Digit. Library Technol. Project, Stanford, CA, USA, Tech. Rep., 2009.
[17] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "SemEval-2016 task 4: Sentiment analysis in Twitter," in *Proc. 10th Int. Workshop Semantic Eval. (Semeval)*, 2016, pp. 1–18.
[18] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge, "Twitter polarity classification with label propagation over lexical links and the follower graph," in *Proc. 1st Workshop Unsupervised Learn. NLP (EMNLP)*, Stroudsburg, PA, USA, 2011, pp. 53–63.
[19] R. Blanco *et al.*, "Repeatable and reliable search system evaluation using crowdsourcing," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, New York, NY, USA, 2011, pp. 923–932.
[20] S. Nowak and S. Rüger, "How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation," in *Proc. Int. Conf. Multimedia Inf. Retr. (MIR)*, New York, NY, USA, 2010, pp. 557–566.
[21] M. Lease, "On quality control and machine learning in crowdsourcing," in *Proc. 11th AAAI Conf. Hum. Comput. (AAAIWS)*, 2011, pp. 97–102.
[22] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Inf. Fusion*, vol. 36, pp. 10–25, Jul. 2017.
[23] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas, "Sentiment analysis leveraging emotions and word embeddings," *Expert Syst. Appl.*, vol. 69, pp. 214–224, Mar. 2017.
[24] N. Liu *et al.*, "Text representation: From vector to tensor," in *Proc. 5th IEEE Int. Conf. Data Mining*, Nov. 2005, p. 4.
[25] L. Derczynski, A. Ritter, S. Clark, and K. Bontcheva, "Twitter part-of-speech tagging for all: Overcoming sparse and noisy data," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process.*, 2013, pp. 198–206.
[26] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2004, pp. 168–177.
[27] F. Å Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," in *Proc. ESWC Workshop Making Sense Microposts' Big Things Come Small Packages*, 2011, pp. 93–98. [Online]. Available: http://arxiv.org/abs/1103.2903
[28] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "NRC Canada: Building the state-of-the-art in sentiment analysis of Tweets," in *Proc. Int. Workshop Semantic Eval. (SemEval)*, Atlanta, GA, USA, 2013, pp. 321–327.
[29] S. Kiritchenko, X. Zhu, and S. M. Mohammad, "Sentiment analysis of short informal texts," *J. Artif. Intell. Res.*, vol. 50, pp. 723–762, Aug. 2014.
[30] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proc. Conf. Hum. Lang. Technol. Empirical Methods Natural Lang. Process.*, 2005, pp. 347–354.

[31] O. Kolchyna, T. T. P. Souza, P. C. Treleaven, and T. Aste. (2015). "Twitter sentiment analysis: Lexicon method, machine learning method and their combination." [Online]. Available: https://arxiv.org/abs/1507.00955
[32] A. Muhammad, N. Wiratunga, R. Lothian, and R. Glassey, "Domain-based lexicon enhancement for sentiment analysis," in *Proc. SGAI Int. Conf. Artif. Intell.*, 2013, pp. 7–18.
[33] J. Bross and H. Ehrig, "Automatic construction of domain and aspect specific sentiment lexicons for customer review mining," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, New York, NY, USA, 2013, pp. 1077–1086.
[34] W. Du, S. Tan, X. Cheng, and X. Yun, "Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining (WSDM)*, New York, NY, USA, 2010, pp. 111–120.
[35] K. Labille, S. Alfarhood, and S. Gauch, "Estimating sentiment via probability and information theory," in *Proc. 8th Int. Joint Conf. Knowl. Discovery, Knowl. Eng., Knowl. Manage. (KDIR)*, 2016, pp. 121–129.
[36] Z. Wang, X. Sun, D. Zhang, and X. Li, "An optimal SVM-based text classification algorithm," in *Proc. Int. Conf. Mach. Learn.*, Aug. 2006, pp. 1378–1381.
[37] B. Pang, L. Lillian, and V. Shivakumar, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. ACL Conf. Empirical Methods Natural Lang. Process.*, vol. 10, Jul. 2002, pp. 79–86.
[38] B. Liu, E. Blasch, Y. Chen, D. Shen, and G. Chen, "Scalable sentiment classification for big data analysis using Naïve Bayes classifier," in *Proc. IEEE Int. Conf. Big Data*, Oct. 2013, pp. 99–104.
[39] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2009.
[40] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[41] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 112. New York, NY, USA: Springer, 2013.
[42] M. Robnik-Šikonja, "Improving random forests," in *Machine Learning: ECML*. Berlin, Germany: Springer, 2004, pp. 359–370.

SAMAH ALOUFI received the M.Sc. degree in computer science from the University of Ottawa, Ottawa, Canada, where she is currently pursuing the Ph.D. degree in computer science. Her research interests include social multimedia mining, and social multimedia retrieval and recommendation.

ABDULMOTALEB EL SADDIK (M'01–SM'04–F'09) is currently a Distinguished University Professor and the University Research Chair of the School of Electrical Engineering and Computer Science, University of Ottawa. He has supervised more than 120 researchers. He has received research grants and contracts totaling over $18 M. He has authored or co-authored 10 books and over 550 publications, and has chaired over 50 conferences and workshop. His research focus is on the establishment of digital twins using AI, AR/VR, and tactile Internet that allows people to interact in real-time with one another as well as with their digital representation. He is an ACM Distinguished Scientist, a Fellow of the Engineering Institute of Canada, and a Fellow of the Canadian Academy of Engineers. He has received several international awards, including the IEEE I&M Technical Achievement Award, the IEEE Canada C. C. Gotlieb (Computer) Medal, and the A.G.L. McNaughton GoldMedal for important contributions in the field of computer engineering and science.