Soccer Fans Sentiment Through The Eye of Big Data:The UEFA Champions League as a Case Study

Samah Aloufi*†, Fatimah Alzamzami*, Mohamad Hoda*, and Abdulmotaleb El Saddik*
*Multimedia Computing Research Laboratory (MCRLab), University of Ottawa, Ottawa, Canada
†College of Computer Science and Engineering (CCSE), Taibah University, Saudi Arabia
Email: saloufi@mcrlab.uottawa.ca,{falza094, mhoda053, elsaddik}@uottawa.ca

***Abstract*—The proliferation of online soccer contents generated by fans and teams has been widely leveraged in the area of sentiment analysis research. This work proposes domain-specific approach for understanding sentiment expressed in soccer conversations. We introduce a soccer-specific lexicon that we leverage in building sentiment model trained on soccer dataset. Our results show the effectiveness of the proposed approach in recognizing fans emotion during soccer events.**

***Keywords*-Soccer; Football; Sentiment Analysis; Machine learning; Big data;**

I. Introduction

Soccer is, by far, the most popular sport in the world among 8000 different types. With 3.5 billion fans [1] and approximately 265 million players [2] around the world, soccer's success in gathering the attention of the public and the media is notable. In 2014, Twitter reported 672 million tweets related to the FIFA World Cup which took place in that year [3]. A block of 18,027,966 tweets relating to the English Premier League Soccer, were collected from Twitter in 2016 [1]. This proves that Twitter has become a venue for soccer fans to discuss everything about their favorite teams and players' performances throughout the season.

Soccer related conversations contain strong emotions shared by fans [2]; hence, they are very well suited for sentiment knowledge discovery [3]. Sentiment analysis has been proven to be effective in predicting opinions and feelings of the population in many real-world cases [4]. In this work, we focus on leveraging sentiment analysis techniques to capture the overall feelings of fans conversations related to the UEFA Champions League (CL) Soccer 2016/2017 championship, and how these feelings change over time. We use a lexicon-based approach combined with a machine learning approach in order to build our sentiment classifier. We utilize a Support Vector Machine (SVM) algorithm trained on a soccer-specific dataset, for providing polarity sentiment (i.e. positive, negative, or neutral) on soccer-related conversations. We then use the extracted sentiment to track the evolution of fans emotion over time. We are particularly interested in tracking the sentiment before, during, and after the games of CL 2016/2017.

The quality of classification performance is directly dependent on determining a good set of features. This is especially true for lexical features, since one word or sentence may inflict different sentiments within different domains [5]. Soccer fans usually have sport talks on social media. These soccer talks might sound like a new language for non fans, especially if combined with slang terms. From a sentiment analysis perspective, using standard sentiment lexicons with soccer talks could lead to learning confusion. For instance, "That long bomb was sick!" indicates a positive sentiment in the soccer domain even though the words "bomb" and "sick" are associated with negative sentiment in general context. To overcome this challenge, we propose to generate a soccer-specific sentiment lexicon that allows us to improve the performance of our sentiment classification. To the best of our knowledge, this is the first work that generates a domain-specific lexicon for soccer sentiment classification.

This paper presents two specific contributions: First, a soccer-specific sentiment model that is capable of recognizing emotional polarity expressed in soccer-language conversations. Second, a soccer-specific sentiment lexicon generated using the FIFA World Cup 2014 dataset.

The rest of the paper is organized as follows: related works are discussed in Section II. In Section III, we describe our sentiment analysis approach. Section IV provides details on experiments and discusses results. In Section V, we provide data analysis on few cases related to CL 2016/2017 soccer event. Finally, Section VI summarizes our findings and possible future work.

II. Related Work

There have been several studies which examine the relationship between sentiment analysis and the prediction of game outcomes [6] and economic profits [7]. Other studies focused on using sentiment analysis techniques to collect the overall opinions expressed by soccer fans through social media. While Barnaghi et al. [3] used FIFA World Cup 2014 tweets to find a correlation between the sentiments expressed in the tweets and events of interest which occurred during

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

0-7695-6354-6/18/$31.00 ©2018 IEEE
DOI 10.1109/MIPR.2018.00058

games, Simeon and Hilderma [8] focused on the overall opinions of the Twitter audience during the matches of FIFA World Cup 2014.

Lexicon-based and machine learning (ML) approaches are widely used in sentiment classification [9]. Authors in [10], [11], [12] adopted a machine learning approach while analyzing FIFA tweets. Barnaghi et al.[10] used external lexicons to improve the ability of Bayesian Logistic Regression (BLR) classifier to detect a higher level of sentiment subjectivity. Support Vector Machine (SVM) algorithm was utilized in [11] and has proven its robustness against Naive-Bayes classifier on the FIFA tweets [12].

One of the most significant steps in sentiment classification is feature extraction. This is due to two reasons: (1) high dimensionality of opinionated text [13], (2) domain considerations, e.g. one word/sentence may reflect different opinions or emotions in different domains, or special slang might be used in a specific domain[9]. Therefore, selecting good indicator features will have a positive impact on the performance of sentiment classifiers. Combining different feature sets would contribute to the improvement of the classification learning [2], [6]. When the sentiment classifier in [10] was trained (i.e. on FIFA World Cup 2014 tweets) based on both Uni-gram and Bi-gram features, it yielded a better performance with an accuracy rate of (74.84%) in comparison to when features were individually applied, resulting in accuracies of (71.35% for Uni-gram, 67.44% for Bi-gram) [3]. Although tweets have a very short text, Uni-gram provided better results on FIFA tweets for sentiment classification [3], [10]. Moreover, results in [14] suggested using a bag of Uni-gram features for topic-based sentiment classification with more than two classes. The choice of lexicons as features for sentiment classification is important; as the classification performance could be compromised if chosen improperly. For instance, using the EmotionWord lexicon to train a sentiment classifier on a sarcastic tweets set actually jeopardizes its performance [5]. Using a combination of lexicons in [11] has shown to be very effective on the FIFA dataset. This better performance on the FIFA dataset could be due to a high level of emotions and opinions contained in fans' tweets [2]

Another important aspect of sentiment analysis is data annotation. Manual annotation of tweets is a very tedious task [15]. To overcome this challenge, researchers in [15] suggested using noisy labeling (i.e. using emoticons, sentiment lexicons, etc.) for automatic sentiment annotation. Authors in [7], [8] used 'off-the-shelf' sentiment analysis libraries (i.e. Sentiment140 [4], Indico APIs [5]) to automatically extract the sentiments of tweets. The sentiment of the tweets was then used to solve event prediction problems [7], or to capture the overall opinions of soccer fans during games [8].

Gratch et al. [2] proposed to train a sentiment classifier on a manually-annotated general dataset and then use the model to predict the sentiment of tweets related to FIFA World Cup 2014. In our work, we follow a similar approach except that we train our proposed model on a domain-specific automatically-labeled tweet dataset.

III. Sentiment Analysis Method

Features extraction and classifier learning are the main components in sentiment classification. Once features sets are extracted, we train our classifier using Linear Suport Vector Machine (SVM). It proves its robust performance in a broad range of applications including text categorization [16]. In the following sections, we provide details of the dataset used for training our model. Then we describe the features adopted in our work.

A. Training Dataset Collection

In order to train our sentiment classifier on soccer specific dataset, we have used the FIFA World Cup 2014 dataset. The FIFA dataset was collected by [3], [10] during the period of June 6th and July 14th 2014 from Twitter. Each tweet was automatically labeled by polarity using Aylien API [6]. The provided list of tweets only included the tweet ids and the polarity labels; thus, we have used Twitter Search API to retrieve the tweet information. Consequently, we have obtained a total of 440,917 tweets. The distribution of tweets over the classes is as follows: 192,717 positive, 133,158 negative, and 115,042 neutral tweets.

B. Features

We have utilized a variety of features, such as Bag-of-words (BOW) which is typically used in sentiment analysis. Since soccer conversations contain high level of emotional reactions [2], we have exploited some existing general lexicons. Moreover, we have developed a new domain-specific sentiment lexicon, which we refer to as FIFA Soccer-Sentiment dictionary.

1) Bag-of-words Features: This is a standard textual feature that is widely used in text analysis. Each instance in the dataset is represented as a feature vector of length d. Each element in the vector corresponds to a term in a pre-defined vocabulary. To generate the vocabulary, we consider two schemes: Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF). Before selecting the vocabulary, we have applied essential Natural Language Preprocessing (NLP) steps such as removing the stop words, the URLs, and the mentions. Also, we have lowercased all the words, normalized elongated words, and applied the Word-Net Lemmatizer using the NLTK toolkit [7]. We have experimented with two different models of BOW features: Uni-gram and Bi-gram. For both sets, we include words that

[4]http://help.sentiment140.com/for-students
[5]https://pypi.python.org/pypi/IndicoIo/0.4.7
[6]https://aylien.com/text-api/
[7]http://www.nltk.org/

occur in at least n tweets within our training dataset. We refer to this process as min-tf in our experiments, where n is 5, 10, 15, 20, and 25 as shown in Tables I and II.

2) Existing Lexicons and Linguistic Features: Many lexicons have been created for sentiment and opinion analysis using different resources. The coverage of single lexicon might be limited; therefore we select four existing lexicon for more diverse exploration of sentiment words as follows: (1) Bing Liu's Opinion Lexicon (OL) [17], (2) AFINN-111 lexicon (AFFIN) [18], (3) NRC Hashtag Sentiment lexicon (NRC) [19], [20], (4), Emoticons and slang lexicon (EMO) [21].

Linguistic signals like ALL-CAPS might emphasis strong emotion in messages. Therefore, we consider four linguistic features: 1) ALL-Caps: number of words in upper case, 2) count of exclamation marks, 3) number of question marks, and 4) the number of combinations of question and exclamation marks. These features were used in [22], [23]. For each lexicon, we count the occurrences of positive and negative terms. Similarly, we count the times each linguistic feature appears in a tweet.

3) Proposed FIFA Lexicon: We argue that fans speak in soccer dialects when interacting with each other on social media. These dialects consist of some vocabularies that might not exist in the standard sentimental lexicons. Even when they intersect, some terms and phrases might refer to different sentiment. For example, words "bomb" and "sick" indicate negative sentiment in the general settings. In soccer context, the sentence "That long bomb was sick!" expresses positive sentiment. Accordingly, we propose to develop a soccer-based sentiment lexicon, which we generated using the training FIFA World Cup 2014 dataset (Section III-A). We follow the steps proposed in [21] to create our lexicon: (1) we follow the same NLP preprocessing steps used in Section III-B1 in order to create a set of vocabulary from both negative and positive tweets. Then, we calculate the number of occurrences of each word in positive and negative tweets within our training dataset. The positive score of a word d_i is then calculated using equation 1 :

$$pos\_score(d_i) = \frac{freq(d_i, positive)}{freq(d_i, positive) + freq(d_i, negative)} \quad (1)$$

where $freq(d_i, positive)$ is the number of times a word d_i occur in positive tweets, and $freq(d_i, negative)$ is the occurrence of d_i in negative tweets. The negative score for word d_i is calculated in a similar way.

The determination of a polarity of a word d_i is made based on its positive score. The word is considered positive if its positive score is above 0.6. All the words with positive scores less than 0.4 are considered negative. Our final lexicon, FIFA soccer-sentiment lexicon, generates a list of 4,784 words: 2,319 of which are labeled as positive and 2,465 negative words.

We believe that using soccer-based lexical features would support our sentiment classifier to adapt to soccer-language tweets. Our results show the effectiveness of the proposed approach for the purpose of this study.

IV. Experiments and Results

The main purpose of the experiments is to evaluate the performance of the proposed model in detecting sentiment of fans reactions during soccer events. To evaluate our model, we use accuracy and F-score metrics. Note that the training is conducted on the FIFA tweets dataset, which was randomly divided into 60%-40% for training and testing, respectively.

The results of BOW features are illustrated in Tables I and II for Uni-gram and Bi-gram models, respectively. When comparing the results in Tables I and II, we observe a significant decrease of the algorithm's performance using Bi-gram model. For both models, including more words in the vocabulary set provides better sentiment results. The Uni-gram model achieves the best accuracy when min-tf was the minimum. From the results, we can see that TF and TF-IDF approaches have achieved similar results.

Table I
Performance of Uni-gram model on classifying sentiment of FIFA tweets

	Accuracy		F-Positive		F-Neutral		F-Negative	
min-tf	TF	TF-IDF	TF	TF-IDF	TF	TF-IDF	TF	TF-IDF
5	**0.839**	**0.840**	0.87	0.87	0.80	0.80	0.84	0.84
10	0.838	0.838	0.87	0.87	0.80	0.79	0.84	0.84
15	0.836	0.835	0.86	0.86	0.79	0.79	0.83	0.83

Table II
Performance of Bi-gram model on classifying sentiment of FIFA tweets

	Accuracy		F-Positive		F-Neutral		F-Negative	
min-tf	TF	TF-IDF	TF	TF-IDF	TF	TF-IDF	TF	TF-IDF
15	**0.654**	**0.656**	0.70	0.70	0.62	0.62	0.62	0.62
20	0.641	0.642	0.69	0.69	0.61	0.61	0.61	0.61
25	0.628	0.631	0.68	0.68	0.60	0.60	0.55	0.60

The experimental results of different lexicons and combination of them are shown in Table III. For individual lexicons, NRC's performance is the worst compared to the others. Limited numbers of emotional hashtags (i.e. not vocabulary) in our dataset have affected the performance of the NRC lexicon. The FIFA lexicon has achieved comparable performance results to the OL lexicon when used individually. Yet, adding FIFA lexicon to the combination of three domain-free lexicons has boosted the accuracy from being 64% to become 68%, as shown in TableIII. This proves the importance of using domain-specific lexicons for sentiment classification.

Table III
COMPARISON OF PERFORMANCE BETWEEN DOMAIN-FREE AND SOCCER-SPECIFIC LEXICONS

Features	Accuracy	F-Positive	F-Neutral	F-Negative
NRC	0.556	0.65	0.32	0.56
OL	0.640	0.69	0.57	0.67
OL+AFINN	0.649	0.66	0.65	0.65
FIFA Lexicon	0.636	0.71	0.35	0.70
EMO+OL+AFINN	0.649	0.70	0.56	0.66
EMO+OL+AFINN+FIFA	0.680	0.73	0.55	0.72
EMO+OL+AFINN+Lingiustics	0.656	0.71	0.56	0.66
EMO+OL+AFINN+Lingiustics+FIFA	0.684	0.74	0.54	0.72
BOW+Lexicons+lingiuistics	0.850	0.87	0.81	0.84

Besides exploring the performance of BOW and lexicons features individually, we have examined the impact of combining multi feature sets on the performance of the classification model. The fusion of BOW, lexicons, and linguistic features has achieved slightly better performance at 85%, than relying solely on a single feature with accuracy of 84%.

V. SENTIMENT ANALYSIS OF CHAMPIONS LEAGUE

Soccer game fans' sentiment are manifested and expressed through their shared tweets during a soccer event. These tweets sequentially reflect the changes and evolution in the fans' sentiment depending on the events of the game, e.g, goal scoring, penalties, etc, as they watch the game. Aggregating the sentiment conveyed through these tweets draw a picture of the feelings that fans are experiencing during a specific soccer event. In our work, we focus on analyzing fans' sentiment during the CL 2016/2017, by utilizing our trained model. In the following sections, we first present the data collection procedure, and then the sentiment analysis is discussed.

A. CL Dataset Collection

To collect the data from Twitter, we have considered the official hashtag '#Championsleague' as a seed to retrieve tweets related to this event. The Twitter Search API has been used for the period of June 1st 2016 to June 15th 2017, which is the duration of the targeted event. In this step, we have obtained 380,579 tweets in multi-languages. To enrich our data coverage of CL event, we have ranked the hashtags that appear in our set based on their occurrence in tweets. Then, the top 15 ranked hashtags have been selected to retrieve more data from Twitter in the same period of time. As a result, we have collected 2,811,833 tweets. From our dataset, 37% of tweets are in English, followed by 25% of tweets in Spanish. In this work, we have considered only English tweets for sentiment analysis. To ensure data quality, we have filtered out duplicated tweets, which left us with 819,848 English tweets. In addition, for the purpose of analysing teams and fan engagement, we have collected English tweets from the teams' official accounts during the CL 2016/2017 season. In total, we have collected 171,121 tweets for all teams participated in CL 2016/2017.

B. Sentiment Analysis During the Season

Figure 1 shows the sentiment results of Champions league divided by month. The aggregation of tweet sentiment during a month can reflect the sentiment of people during matches which occurred within this period of time. For example, in March 2017, 8 games were played as part of round of 16 stage. Most of the tweets posted during March 2017 reflect the negative emotion expressed by the teams' fans. Barcelona's defeat of Paris Saint-Germain (PSG) 6-0 on March 8th provoked controversy among fans. PSG fans were elated after their victory by 4 goals on the first leg against Barcelona. They were equally devastated by the subsequent loss since PSG then failed to qualify to Quarter-Finals. On other side, Barcelona fans went from being crestfallen at the possibility of elimination to jubilant when they did, in fact, advance to the Quarter-Finals. In addition, Bayern Munich defeated Arsenal 5-1, which resulted in having more negative tweets than positive ones. Arsenal tweets were shared for days after the game and generated more interactions, as fans were comparing the Arsenal loss to PSG's defeat. The negative impact of Arsenal-Bayern match results could be related to the higher number of tweets mentioning Arsenal, in that it generated 15,241 tweets, compared to only 5,438 tweets for Bayern. This could be due to the fact we only considered English tweets in our study. Also, the higher number of negative tweets during this month could be explained by the tie result of Atletico-Leverkusen match, which lead to Leverkusen failing to be qualified to Quarter-Finals.

Approximately 70,370 tweets related to the Quarter-Finals were collected. Figure 1 shows that the positive sentiment was the dominant during the month of April followed by the negative sentiment. The high level of negative sentiment during this month could be correlated to the heavy loss of Barcelona against Juventus (0-3 for Juventus) on April 11, and the tie game between the same teams on April 19. It is important to mention that Barcelona has the highest number of tweets among all other teams during the Quarter-Finals. The team has 21.57% of the tweets followed by Real Madrid, Juventus, and Leicester (18.22%, 17.48%, and 15.62%, respectively). Hence, the sad feeling of Barcelona fans are obvious when interpreting the results. The negativity of sentiment during the Quarter-Finals could also be attributed to the fact that both Barcelona and Leicester did not qualify for the Semi-Finals. On the other hand, Fans of Real Madrid and Juventus expressed mostly positive sentiment, since Real Madrid won both games and Juventus marked a win (3 goals for 0) against Barcelona, and, most importantly, qualified to play in the Semi-Finals.

We have observed that there is a relationship between fan engagements and the social activities of teams. Barcelona and Juventus, for example, use strategies to drive their fans engagements. They publish news and live updates during matches and respond to the fans' posts through their official accounts. Juventus decided to trigger the emotional aspect of their fans by creating the hashtag #ItsTime for the Final stage. This strategy was a success with the fans and the hashtag #ItsTime was used more than their main hashtag #finoallafine (99 times for #finoallafine, 127 times for #ItsTime in our dataset).

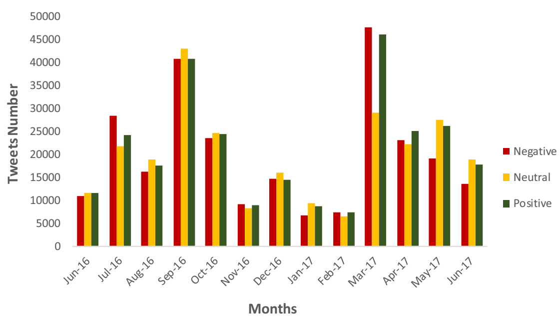

Figure 1. Overall sentiment during CL season

C. Sentiment Analysis of The Final Game

We considered all the tweets related to the final match between Juventus (Juve) and Real Madrid (RM). To extract the tweets, we used the hashtags of the teams and the ones related to the game. Our objective is to examine whether an occurrence of major events during the games would drive fans to engage more than usual on social media. We assume that more social conversations would convey more emotions.

Tracking the sentiment of fan talks throughout the game has revealed interesting insights. Figure 2 illustrates the feeling of fans before, during, and after the game. The time window was set to be 2 hours prior to the game starts and 2 hours after game end. Fans of both teams shows positive sentiment leading up to the game kick off. A sample of related tweets shows a lot of support and cheering attitude, hence, the positive sentiment. During the game Juventus fans were not happy about the team performance; whereas supporters of Real Madrid showed a very positive sentiment. This is not surprising since Real Madrid stunned Juventus 4-1 in the Final. When ignoring objective tweets, 52% of Juventus tweets contained negative emotion while positive sentiment from Real Madrid's covered 58% of the overall subjective tweets. Post-game sentiment continues the same trend as those logged over the duration of play. Fans of Real Madrid expressed happiness in 74% of the related subjective tweets.

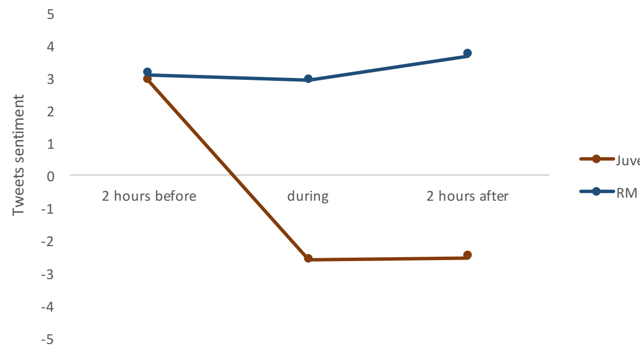

Figure 2. Sentiment changes prior, during, and post game in CL Finals

More investigations were conducted on the sentiment during the game. We are particularly interested in examining sentiment changes displayed during occurrences of major events such as scoring goals. We divided the gametime tweets into 7 parts each representing a 15-minute portion of the game.

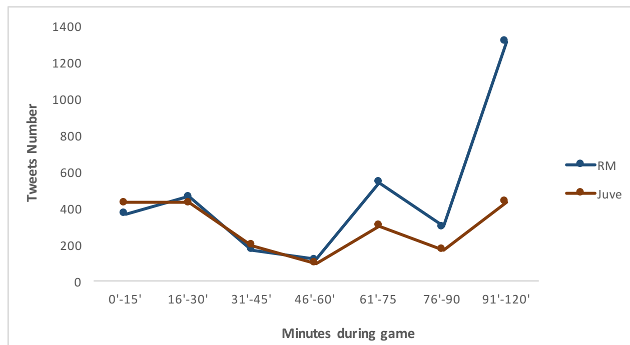

Figure 3. Relationship between fans activities and occurrence of goal-scoring events

We have observed a sudden increase in fans activities when a goal is scored, as shown in Figure 3. During the second 15 minutes of the game (16'-30'), Real Madrid scored its first goal. This explains a sudden increase in Real Madrid tweets during the same time point. Both teams were associated to positive sentiment (i.e. more than 51% convey positiveness) as a result of a 1-1 tie in the first half of the game, as illustrated in Figure 4. Real Madrid scored 2 goals between the 60'-65' minute point of the game. Again, this has resulted in another sudden increase in the volume of tweets by fans of both teams. Obviously, Juventus fans expressed negativity (i.e. over 55% of their tweets is negative) against the game's results, including a contrary opinion to the fans of Real Madrid in the second half of the game. The sentiment results in the second half depicted on Figure 4, reflect a high level of positiveness in their conversations (more than 54% of their tweets is positive) that dramatically increased (Figure 3) after Asensio

(RM player) scored the 4th goal on the 90th minute. Note that Fans usually mention both teams in their tweets which, in turn, affect the overall true sentiment representation for each team.

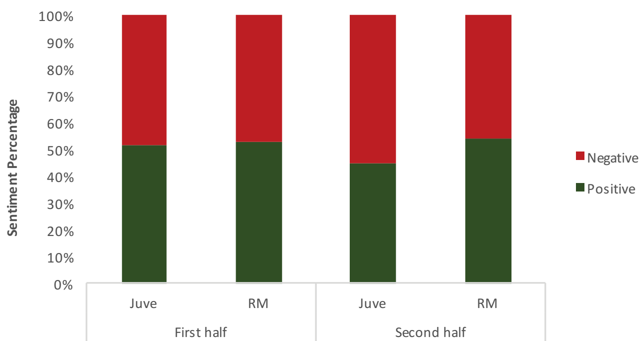

Figure 4. Sentiment changes between the first and second halves of the game in CL Finals

VI. Conclusion

This work describes the process of automatically creating domain-specific lexicon from a soccer dataset. Our results demonstrates the effectiveness of our proposed approach in recognizing sentiments expressed in soccer conversations prior, during, and post games. Moreover, we found a relationship between fans activities and the occurrence of goal-scoring event. We found that occurrences of goals generate sparsity of sentiment between positive for the scoring team and negative for the opponent team. The sentiment in this work was conveyed using only English tweets. Thus for a future improvement we will consider multi-languages tweets to cover a broader sentiment representation. Also, we consider improving our soccer lexicon by integrating domain expert knowledge.

References

[1] R. P. Schumaker, A. T. Jarmoszko, and C. S. Labedz, "Predicting wins and spread in the premier league using a sentiment analysis of twitter," *Decision Support Systems*, vol. 88, pp. 76–84, 2016.

[2] 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 *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 898–903.

[3] P. Barnaghi, P. Ghaffari, and J. G. Breslin, "Text analysis and sentiment polarity on fifa world cup 2014 tweets," in *Conference ACM SIGKDD*, vol. 15, 2015, pp. 10–13.

[4] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series." *ICWSM*, vol. 11, no. 122-129, pp. 1–2, 2010.

[5] N. Malandrakis, M. Falcone, C. Vaz, J. Bisogni, A. Potamianos, and S. Narayanan, "Sail: Sentiment analysis using semantic similarity and contrast," *SemEval 2014*, p. 512, 2014.

[6] S. Sinha, C. Dyer, K. Gimpel, and N. A. Smith, "Predicting the nfl using twitter," *arXiv preprint arXiv:1310.6998*, 2013.

[7] L. Le, E. Ferrara, and A. Flammini, "On predictability of rare events leveraging social media: a machine learning perspective," in *Proceedings of the 2015 ACM on Conference on Online Social Networks*. ACM, 2015, pp. 3–13.

[8] C. Simeon and R. Hilderman, "Evaluating the effectiveness of hashtags as predictors of the sentiment of tweets," in *International Conference on Discovery Science*. Springer, 2015, pp. 251–265.

[9] B. Pang, L. Lee *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[10] P. Barnaghi, P. Ghaffari, and J. G. Breslin, "Opinion mining and sentiment polarity on twitter and correlation between events and sentiment," in *Big Data Computing Service and Applications (BigDataService), 2016 IEEE Second International Conference on*. IEEE, 2016, pp. 52–57.

[11] C. Simeon, H. J. Hamilton, and R. J. Hilderman, "Word segmentation algorithms with lexical resources for hashtag classification," in *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 743–751.

[12] A. L. Firmino Alves, C. d. S. Baptista, A. A. Firmino, M. G. d. Oliveira, and A. C. d. Paiva, "A comparison of svm versus naive-bayes techniques for sentiment analysis in tweets: a case study with the 2013 fifa confederations cup," in *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*. ACM, 2014, pp. 123–130.

[13] A. Sharma and S. Dey, "Performance investigation of feature selection methods and sentiment lexicons for sentiment analysis," *IJCA Special Issue on Advanced Computing and Communication Technologies for HPC Applications*, vol. 3, pp. 15–20, 2012.

[14] 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*. Association for Computational Linguistics, 2002, pp. 79–86.

[15] N. N. Bora, "Summarizing public opinions in tweets," *International Journal of Computational Linguistics and Applications*, vol. 3, no. 1, pp. 41–55, 2012.

[16] B. Kwolek and S. Sako, "Learning siamese features for finger spelling recognition," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2017, pp. 225–236.

[17] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '04. New York, NY, USA: ACM, 2004, pp. 168–177.

[18] F. Å. Nielsen, "A new ANEW: evaluation of a word list for sentiment analysis in microblogs," *ESWC2011 Workshop*, vol. abs/1103.2903, pp. 93–98, 2011.

[19] S. Kiritchenko, X. Zhu, and S. M. Mohammad, "Sentiment analysis of short informal texts," *Journal of Artificial Intelligence Research*, vol. 50, pp. 723–762, 2014.

[20] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "Nrccanada: Building the state-of-the-art in sentiment analysis of tweets," in *Proceedings of the International Workshop on Semantic Evaluation, SemEval13*. Citeseer, 2013.

[21] O. Kolchyna, T. T. P. Souza, P. C. Treleaven, and T. Aste, "Twitter sentiment analysis: Lexicon method, machine learning method and their combination," *http://arxiv.org/abs/1507.00955*, vol. abs/1507.00955, 2015.

[22] Z. Li, S. Zhu, H. Hong, Y. Li, and A. El Saddik, "City digital pulse: a cloud based heterogeneous data analysis platform," *Multimedia Tools and Applications*, vol. 76, no. 8, pp. 10 893–10 916, Apr 2017.

[23] N. Plotnikova, M. Kohl, K. Volkert, S. Evert, A. Lerner, N. Dykes, and H. Ermer, "Klueless: Polarity classification and association." in *SemEval@ NAACL-HLT*, 2015, pp. 619–625.