2014 FIFA World Cup: An Initial Analysis of Collective Sentiments in Twitter

Rubens Pessoa[1], **Jonathas Magalhães**[12],
Marlos Silva[12], **Henrique Pacca**[1], **Evandro Costa**[1]

[1]Intelligent, Personalized and Social Technologies Group
Federal University of Alagoas - Maceió, Brazil.

[2]Artificial Intelligence Laboratory
Federal University of Campina Grande - Campina Grande, Brazil

{rpbf,evandro}@ic.ufal.br, jonathas@copin.ufcg.edu.br

marlos.tacio@gmail.com, henrique.luna@pq.cnpq.br

Abstract. *Social analysis in Twitter has become a very interesting research scenario. Twitter's popularity and the millions of spontaneous interactions between its users turned the social network into a rich data source. A possible use of this data is to measure collective sentiments related to events. In this short paper, we propose an analysis of the content shared in Twitter during the event 2014 FIFA World Cup Brazil™. The sentiment classifier utilized in this work is the SentiStrength tool. The steps of this works are Composing the corpus, Preprocessing, Classification and Temporal Analysis. Finally, we propose some preliminary results that indicates the distribution of sentiments across the corpus.*

1. Introduction

The scenario that emerged from Web 2.0 brought a new context for users to interact with the internet. Leaving aside a passive way of browsing the web to contribute actively among other users to the generation of content. This data is generated by users in several manners, but mostly via social networks or microblogging platforms. This change of perspective created what we know today as big data and this massive generation of content also created several challenges for the academia.

It's important to notice that this data is created in a spontaneous way and represents the user opinion. In this sense, sentiment analysis has emerged proposing to detect the opinion of the crowd about some theme [Pang and Lee 2008]. The task is very challenging, but very useful. For example, an event organizer wants to know what people are saying about a given event. Potential consumers also want to know the repercussion of a given event before actually going. According to [Liu and Zhang 2012], Sentiment Analysis, also known in the literature as Opinion Mining, is the computational study of people's opinion, emotions and attitudes related to entities, services, events, governments, individuals and topics.

Microblogging became one of the main communication tools available on the Web 2.0. Daily there are millions of interactions between its users. Those interactions generally reflect what the users think about their lifes or reflect their opinions related to a

discussion. A broadly used microblogging platform is Twitter[1] , which consists on a website where users can write what they are thinking in a message of 140 characters. These messages are called tweets. Given its characteristics, Twitter turned out to be an excelent data source of public opinions related to daily life [Pak and Paroubek 2010].

In this paper, we propose an analysis of the collective sentiments related to the 2014 FIFA World Cup. The steps of this investigation are: Composing the corpus, Preprocessing, Classification and Temporal Analysis. In the first step, we gathered tweets related to the 2014 FIFA World Cup that were posted during the event. These tweets are in portuguese and english languages, but in this work only tweets in english were analysed. The sentiment classifier utilized in this work is the SentiStrength tool. Finally, we propose some preliminary results that indicates the distribution of sentiments across the corpus.

The remainder of this paper is structured as follows. Section 2 presents some related works. Section 3 describes The Analysis System. Section 4 presents some preliminary results. Section 5 concludes with proposal for future work.

2. Related Work

The advent of social media has enabled the dissemination of information within users in social networks. According to [Twitter 2014], 500 million tweets are shared by its users per day. A large part of these tweets contains opinionated or emotional content related to events or entities. Hence, analysing Twitter trending topics turned into a scenario that has been widely researched by the academia. Works on this area of study covers how to understand the content available and how it can be used.

[Sakaki et al. 2010] propose an algorithm for real-time monitoring and detecting the occurrence of natural disasters simply using content shared within users in Twitter. In this work, semantic analysis is used to select useful tweets and Support Vector Machine (SVM) is used in order to classify the content of tweets into positive or negative.

[Nguyen et al. 2013] investigate tweets related to the royal birth of Prince George of Cambridge in 2013 across different geographic areas during a period of 3 days. In this work, they describe a framework in order to explain the steps of the investigation. The framework is divided into stages from collection of tweets to the graphical visualization of sentiments. This work also indicates that an implementation of faster methods using big data techniques in order to make real-time sentiment analysis of tweets is required.

[Pak and Paroubek 2010] describes an approach for sentiment analysis in Twitter corpus. In order to train a Naïve Bayes classifier to recognize the positiveness of each tweet, a method for collecting two corpus (positive/negative corpus and neutral corpus) is described. These corpus are utilized in order to understand the lexical differences between sentimental and not sentimental messages. Secondly, the corpus is tokenized into n-grams and some indicators of sentiment are perceived among different forms of writing. This work found that using bi-grams provides better accuracy than unigrams or trigrams. Their classifier is able to recognize positive, negative and neutral content.

[Firmino Alves et al. 2014] proposes an accuracy comparison between two machine learning approaches applied to sentiment analysis: Support Vector Machine and

[1] https://twitter.com/

Naïve Bayes. The general idea of this work was to replace POS Tagger for identification of opinionated tweets. The approach utilizes two stages of classification. First, it classifies tweets in opinionated or not-opinionated. Secondly, it classifies according to the feeling carried in each tweet. This work utilized a case study namely 2013 FIFA Confederations Cup.

3. The Analysis System

Figure 1 describes the steps of identifying collective sentiments in Twitter. Despite being instantiated on Twitter in this work, these steps can be used in various domains in order to visualize how people react to facts. The steps are further explained in the next topics.

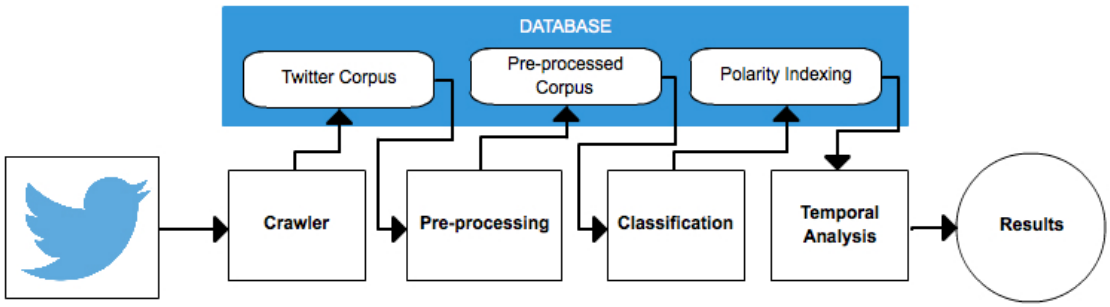

Figure 1. Framework for analysing collective sentiments in Twitter

3.1. Composing the Corpus

Almost 4,2 million tweets in english and portuguese were collected during the month between 12 June and 18 July of 2014 containing words related to the FIFA World Cup 2014. In this work we focus on english language tweets. There are 1.8 million tweets in those conditions, which corresponds to approximately 43% of the database. The Twitter Streaming API[2] was used to collect this data from Twitter servers in the time they were shared. The crawler was created to collect tweets that contains at least one of the following keywords: "World Cup", "World Cup Brazil", "World Cup 2014", "World-Cup", "#WorldCup" and "#WorldCup2014". The data stored in the database was in JSON format. Each JSON is linked to a specific tweet and contains the tweet metadata. Hence, we can store several informations related to each tweet (e.g. timing and local). These characteristics are provided by the Twitter Streaming API. The possibility of collecting these attributes makes it possible to perform temporal analysis and geographical analysis. [Pak and Paroubek 2010] [Hassan et al. 2013].

3.1.1. Pre-processing

After stored in the database, every tweet was pre-processed in order to facilitate the Sentiment Classification step. The pre-processing step consisted in removal of the following: user names, special terms ("RT", "via", "#") and stopwords. Finally, blanks were replaced by "+". Hence, the pre-processed sentence could be used on SentiStrength[3] tool.

[2] https://dev.twitter.com/streaming/overview
[3] http://sentistrength.wlv.ac.uk

3.2. Sentiment Classification

This step classifies tweets according to the sentiment that the tweet loads. We utilized SentiStrength in order to classify the sentiment of tweets. SentiStrength is a sentiment analysis tool that uses a lexical approach that utilizes a list of sentiment-related terms and has rules to deal with contemporary variances of linguistics on web [Thelwall and Cybermetrics].

According to [Nguyen et al. 2013], there is a good correlation related to the results of popular dictionary-based approaches and machine learning approaches when large data are analysed and in the [Gonçalves et al. 2013] work, SentiStrength approach showed to have an acceptable coverage and the second better F-measure, losing only to the Emoticons approach which has a coverage less than 10 percent. Hence, SentiStrength has the best relation between coverage and F-measure.

3.2.1. Tweet Classification

The output of the SentiStrength tool is a tuple composed by maximum positive sentiment score and minimum negative sentiment score of a sentence. The score of positive sentiment varies from 1 to 4 and the score of negative sentiment varies from -1 to -4. For facilitating the classification and retrieval step, both scores are added together. In this sense, let RS represents the sum and PC represents the polarity classification, so the classification step follows the piecewise-defined function:

$$PC(RS) = \begin{cases} positive & \text{if } RS \geq 0 \\ negative & \text{if } RS < 0 \end{cases}. \tag{1}$$

3.2.2. Temporal Analysis

In this step, we propose a sentiment classification for given periods of time. The output is a sentiment that represents the feeling of the majority of tweets posted during the given period. This is the most important and interesting step of all this work. Once we have the sentiment classification of days, weeks and weekends, we will search for correlations between what really happened related to the FIFA World Cup 2014 in those periods of times and the majority sentiment detected. We will identify most frequently used words in order to facilitate the understanding of what people wanted to say in this period.

This step can be utilized in various domains, inclusive in real-time. In this case, techniques of Big Data are necessary in order to deal with massive user generation of data. [Nguyen et al. 2013]

4. Preliminary Results

The twitter corpus collected to make this work possible has nearly 4,2 million tweets related to the 2014 FIFA World Cup. Table 1 presents the quantity of tweets in the dataset per keyword. Figure 2 presents the sentiment distribution of tweets in english. This way of classifying shows the intensity of the feelings contained in tweets. How lower the rating, the more negative feelings and vice versa.

It's important to notice that there are intersections containing two or more keywords in this corpus. The reason of this characteristic may be because Twitter users often use more than one keyword (e.g. hashtags) to expand the reach of their messages to other users. [Kwak et al. 2010]

Table 1. Quantity of tweets per keyword

Keyword	Quantity
"World Cup"	2.243.911
"World Cup Brazil"	10.109
"World Cup 2014"	167.862
"WorldCup"	1.449.019
"#WorldCup"	1.336.617
"#WorldCup2014"	422.528

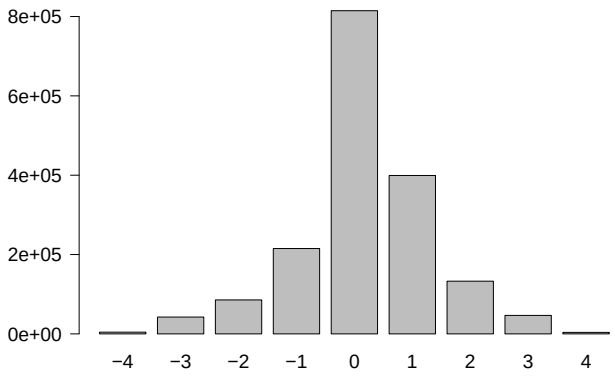

Figure 2. Sentiment Distribution

5. Conclusions and Future Work

This paper described a framework to identify collective sentiments in social networks. The stages of this framework are: Composing the Corpus, Pre-processing, Classification and Temporal Analysis. The classification step utilizes SentiStrength tool in order to identify the feelings carried by every single tweet in the dataset. Hence, Collective Sentiment Score (CSS) can be calculated in the Temporal Analysis stage. The final output is a binary score of feelings (i.e. positive and negative). A case of study, namely 2014 FIFA World Cup Brazil, was proposed in order to validate our work.

Plans for future work are: perform the analysis system proposed in this work through a case of study 2014 FIFA World Cup. Identify correlations between what happened in the 2014 World Cup and the crowd feeling, in addition to identifying most frequently used words in order to facilitate the understanding of what people wanted to

say; accuracy comparison between dictionary-based approach and machine learning approaches implemented [Pak and Paroubek 2010]; avoiding noises in streaming content seeking to increase overall accuracy in prediction algorithms similar to the work proposed by [Hassan et al. 2013]; investigate visualization techniques in order to facilitate the verification of results, similar to demonstrated in [Nguyen et al. 2013].

References

Blog, T. (2014). Insights into the #worldcup conversation on twitter. Access in: 2014/09/10.

Firmino Alves, A. L., Baptista, C. d. S., Firmino, A. A., Oliveira, M. G. a. d., and Paiva, A. C. d. (2014). 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*, WebMedia '14, pages 123–130, New York, NY, USA. ACM.

Gonçalves, P., Araújo, M., Benevenuto, F., and Cha, M. (2013). Comparing and combining sentiment analysis methods. In *Proceedings of the First ACM Conference on Online Social Networks*, COSN '13, pages 27–38, New York, NY, USA. ACM.

Hassan, A., Abbasi, A., and Zeng, D. (2013). Twitter sentiment analysis: A bootstrap ensemble framework. In *Social Computing (SocialCom), 2013 International Conference on*, pages 357–364.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA. ACM.

Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Aggarwal, C. C. and Zhai, C., editors, *Mining Text Data*, pages 415–463. Springer US.

Nguyen, V. D., Varghese, B., and Barker, A. (2013). The royal birth of 2013: Analysing and visualising public sentiment in the uk using twitter. In *Big Data, 2013 IEEE International Conference on*, pages 46–54.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA. ACM.

Thelwall, M. and Cybermetrics. Heart and Soul: Sentiment Strength Detection in the Social Web with SentiStrength 1. 5.

Twitter (2014). About twitter, inc. — about. Access in: 2014/08/15.