A Fine Grain Sentiment Analysis with Semantics in Tweets

Cristóbal Barba González, José García-Nieto, Ismael Navas-Delgado, José F. Aldana-Montes

Lenguajes y Ciencias de la Computación, Universidad de Málaga

***Abstract* — Social networking is nowadays a major source of new information in the world. Microblogging sites like Twitter have millions of active users (320 million active users on Twitter on the 30th September 2015) who share their opinions in real time, generating huge amounts of data. These data are, in most cases, available to any network user. The opinions of Twitter users have become something that companies and other organisations study to see whether or not their users like the products or services they offer. One way to assess opinions on Twitter is classifying the sentiment of the tweets as positive or negative. However, this process is usually done at a coarse grain level and the tweets are classified as positive or negative. However, tweets can be partially positive and negative at the same time, referring to different entities. As a result, general approaches usually classify these tweets as "neutral". In this paper, we propose a semantic analysis of tweets, using Natural Language Processing to classify the sentiment with regards to the entities mentioned in each tweet. We offer a combination of Big Data tools (under the Apache Hadoop framework) and sentiment analysis using RDF graphs supporting the study of the tweet's lexicon. This work has been empirically validated using a sporting event, the 2014 Phillips 66 Big 12 Men's Basketball Championship. The experimental results show a clear correlation between the predicted sentiments with specific events during the championship.**

***Keywords* — Microblogging, Big Data, Sentiment Analysis, Apache Hadoop, MapReduce, Twitter, RDF, Named-Entity Recognition, Linked Data.**

I. Introduction

SOCIAL network tools are becoming an important means of social interaction. In this sense, public messages are being analysed to track user opinions on any relevant aspect. Thus, companies are using this kind of analysis to gain insight into the success of new products and services. In this context, Twitter has grown in popularity in recent years and now reports (30th September 2015) that it has a volume of approximately 500 million tweets sent per day and 320 million active users monthly [1]. Therefore, Twitter is a key source of real time opinions of millions of clients or potential clients and so, a valuable source of information for companies.

The term "Big Data" refers to data that cannot be processed or analysed using traditional techniques. Big Data Analytics enables information retrieval from these data. Many software solutions linked to the Apache Hadoop framework [3] are developed to solve problems encountered through the analysis of social media. The problem of analysing tweet streams is a typical example of the use of the Hadoop ecosystem as its technology can provide solutions to make it feasible.

The automatic interpretation of the results of a Big Data analysis, by exposing their semantics and preserving the context of how they have been produced, are some of the challenges to be addressed when trying to add these results to business processes. In this scenario, the concept of *Smart Data* emerges, which could be defined as *(Def. 1)*: *the result of the process of analysis performed to extract relevant information and knowledge from Big Data, including context information and using a standardized format*. By context, we mean all the relevant metadata needed to interpret the analysis of results. This leads to the enforceability of these results thereby facilitating their interpretation, easy integration with other structured data, integration of the Big Data analysis system with the BI systems, and the interconnection (in a standardised way, at a lower cost and with higher accuracy and reliability) of third party algorithms and services.

This has motivated us to propose here, a novel approach for performing sentiment analysis on tweet streams using Big Data technology, although with the aim of obtaining Smart Data. This approach follows the MapReduce programming model for the analysing of tweets by means of an ontology-based [23] text mining method. The analysis is not limited to just calculating the sentiment value of a given tweet, but also the sentiment of entities mentioned in the text. A domain ontology guides the analysis process, providing sentiment values as RDF graphs [23]. The use of RDF enables the publication of the analysis results as Linked Data and their integrated use with other Linked Data repositories.

In this context, there are no benchmarks for measuring the quality of fine grain sentiment analyses. This led us to select a case in which we could relate measured sentiment values with real life events. Thus, if the sentiment analysis correlates these events we have an empirical validation of the proposed solution. In this paper we present the use of sporting events for this validation. The chosen event is the 2014 Phillips 66 Big 12 Men's Basketball Championship [2]. The Big 12 is a set of sporting events, founded in 1994, including sports such as basketball, baseball, and American football. In the 2014 event, 10 universities from around the United States participated: Iowa, Kansas, Oklahoma, Texas and West Virginia. These universities are Baylor University, Iowa State University, Kansas University, Kansas State University, Oklahoma University, Oklahoma State University, Texas Christian University, Texas Tech University and West Virginia University.

The main contributions of this paper are summarised as follows:

- A MapReduce algorithm for the sentiment analysis of tweets that incorporates a semantic layer to improve the text mining, is proposed for the first time.
- A thorough experimentation of our proposal is carried out from three different viewpoints. First, the analysis of the number of tweets and their relation to the match throughout the championship. Second, the analysis of the relationships between sentiment in the tweets and match scores. Third, the use of linear regression to study the relation between number of tweets and sentiment values.

The remainder of this article is organised as follows. The following section presents background concepts. Section 3 reviews the related

DOI: 10.9781/ijimai.2016.363

literature. In Section 4 we present the problem description. Section 5 details our MapReduce approach. In Section 6, experimental results are presented. Section 7 includes a discussion of results, and finally, Section 8 extracts conclusions from this discussion and details future work.

II. Background Concepts

In this section, we describe the different concepts and tools used in this paper, for the sake of a better understanding.

When we think of Big Data Analytics, one of the main frameworks used to address it is Apache Hadoop (Hadoop) [3]. Hadoop is a software framework for the distributed processing of large data sets across clusters of computers using simple programming models. The core of Hadoop consists of a storage component, known as Hadoop Distributed File System (HDFS), and a processing engine called MapReduce.

HDFS is a Java-based file system that provides scalable and reliable distributed data storage. It has been designed to span large clusters of commodity servers. HDFS is a scalable, fault-tolerant, distributed storage system that works closely with a wide variety of concurrent data access applications, usually coordinated by MapReduce.

MapReduce is an increasingly popular, distributed computing framework for large-scale data processing that is amenable to a variety of data intensive tasks. Users specify serial-only computation in terms of a *map* method and a *reduce* method. The underlying implementation automatically parallelises the computation, offers protection against machine failures and efficiently schedules inter-machine communication [4].

The Natural Language Processing (NLP) is used to analyse the tweets. In this approach, we use GATE (General Architecture for Text Engineering [5]). This suite, developed at the University of Sheffield, is used for entity recognition. The GATE developer module is used for a first syntactic analysis that, in our proposal, is complemented with a semantic (ontology guided) analysis of the tweets. The GATE developer contains a component for information extraction (ANNIE) that determines, from an input text, the different terms that compose it. This component divides the text into simple terms such as punctuation marks, numbers or words. During this process, the component identifies certain types of special rules for English, which enables a more effective division (such as contracted forms like "*don't*" or the Saxon genitive). The division into terms helps us to identify the entities described in the domain ontology (populated with synonyms of the different instances).

SentiStrength [6] is a tool used for the identification of the calculation of the sentiment values. SentiStrength lets us perform a quick test, by inputting a phrase that can even include emoticons, acronyms, etc., classifying it on the positive and negative scale. The positive values (sentiment) range between 1 (not positive) and 5 (extremely positive). The negative values (sentiment) range between -1 (not negative) and -5 (extremely negative). This analysis is done in the context of a recognised entity in a tweet. This way, a tweet can deliver several sentiment values for each different entity.

Sentimental force is specific to the context in which the word tends to be used. SentiStrength employs a machine learning algorithm to optimise the emotional power of words in a sentence, which is incremented or decremented by 1 strength point depending on how the accuracy of the ratings increases. Another issue to consider is that the strength of the emotion of a sequence of words can be altered by the words that precede them in the text. SentiStrength includes a list of words that increase or decrease the excitement of a sequence of subsequent words, in a positive or negative sense. Each word in the list increases the strength of emotion by 1 or 2 points (e.g. very or extremely words) or decreases it by 1 point (for example, the word some). The algorithm also has a list of words that reverse the polarity of the subsequent emotion words, including any amplifier preceding word (e.g., "*very happy*" is a positive force for 5 points, but "*not very happy*" has a negative power of 5 points).

SentiStrength employs a spelling correction algorithm that identifies the standard spelling of words that are misspelled by the inclusion of repeated letters. The algorithm also considers the use of repeated letters commonly used to express emotion or energy in the texts, so before correcting them orthographically it increases the emotion words by 1 point, provided there are at least two additional letters (one repeated letter commonly appears in a misspelling).

The use of emoticons is usual in social networks, so the list of forces increases feelings with a list of emoticons (plus or minus 2 points). In addition, any sentence with an exclamation mark is assigned a positive sentiment at least.

III. Literature Overview

In [7] Barbosa et al. propose a 2-step sentiment detection. The first step targets distinguishing subjective tweets from non-subjective or subjective tweets. The second step further classifies the subjective tweets into positive, negative and neutral. This method is called polarity detection. The authors use polarity predictions from three websites as noisy labels to train a model and use 1000 manually labelled tweets for tuning and another 1000 manually labelled tweets for testing.

Argawl et al. [8] have designed a series of models for classifying tweets into positive, negative or neutral sentiments. Their approach proposes two classification tasks: a binary task and a 3-way task. Thus, they use three types of models for the classification: a unigram model, a feature based model and a tree based model and various combinations of the three. For the tree kernel based model, they designed a new tree representation for tweets. The feature based model uses 100 features and the unigram model uses over 10,000 features. They concluded that combining prior polarity of words with their parts-of-speech tags is the most important part of the classification task. They also point out that the tree kernel based model outperformed the other two.

One of the main features of tweets is also their major drawback i.e., their length. Users have to express their thoughts in just 140 characters. Consequently, the messages must be short so people have to use acronyms, emoticons and other characters that convey special meanings and introduce misspelling. Understanding the meanings of these characters is essential for interpreting the sense of the tweets [6]. In this regard, Kumar et al. [9] propose the use of a dictionary-based method and a corpus to find orientation of verbs and adverbs, thus supplying calculated sentiment to the tweets.

Finally, Monchón et al. [25] propose a new study on the measurement of happiness in Latin America, which shows the possibility of measuring the happiness through the use of social networks, and so it is tremendously simple to calculate via objective and empirical means.

In our proposed method, we aim to go one step beyond these previous approaches by incorporating a semantic model to improve the entity identification in the text mining phase, and hence enhance the sentiment analysis.

IV. Problem Description

The problem faced here is the identification of the sentiment at entity level in a set of tweets. Part of the Big 12 competition is used for testing purposes. This data subset is the tweets of the last three competition days for the Big 12 men's basketball championship.

The main challenge in this work is to identify entities of interest found in tweets. This implies the identification of the list of words related to this event, for instance: teams, players, coaches, referee, and matches. Another problem is the number of related tweets that we can filter from Twitter in almost real time using Twitter API [11]. This requires a continuous process of improvement on the filtering keywords used to reach relevant tweets.

In order to address this problem, we have used Hadoop as the parallelization mechanism since the process fits well in the Map-Reduce methodology.

V. Map Reduce Proposed Approach

In this section, we introduce a methodology that combines the distributed computing framework MapReduce with an ontology-based text mining approach to apply fine grain sentiment analyses.

The sentiment analysis uses SentiStrength. The semantic context introduced by the domain ontology enables the early filtering of the tweets based on relevant entities for the analysis tasks. In the use case presented here, many tweets have been discarded because their content has been determined to be irrelevant for the analysis.

The proposed methodology is illustrated in Fig. *1* and includes the following elements:

- The **domain ontology** to describe the analysis context. This ontology includes not only the structure, but also the population of each term with domain knowledge expressed as ontology instances. This knowledge includes names and synonyms for the entities that we aim to recognise in the tweets.
- The tweet **extraction and analysis algorithms**. This process is done using MapReduce as the programming methodology. The search of new tweets is dynamic and can include new search terms as soon as they are detected as relevant during the analysis process.
- The **NLP analysis** algorithm. This algorithm has been developed using GATE and is guided by the domain ontology.
- The **sentiment calculation** algorithm via SentiStrength, later analysed according to the aggregation level needed.

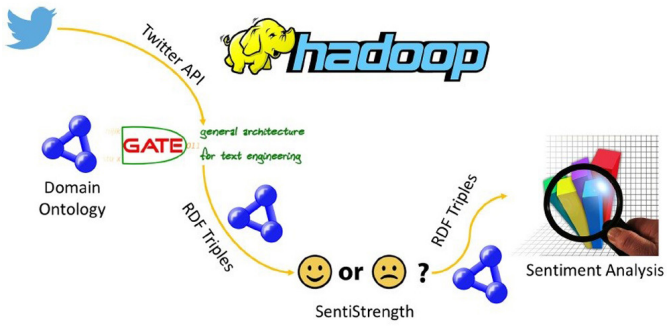


Fig. 1. Methodology for tweet analysis

A. Semantic layer

We have built a full semantic annotation rather than a simple tagging at tweet level. This means that we can identify the location of a term in the text that corresponds to the underlying entity. Ontological proximity disambiguates entities in the text by looking at relationships between entities matched in the tweets. Entities that have a close ontological relationship are deemed to be more likely to be correct. For example, if an analysis of tweets identifies both Texas (university team) and Texas (state) and basketball, then Texas (state) will be disambiguated due to the close ontological relationship with basketball. We have defined an ontology with the concepts of interest for the analysis task. In our use case, these concepts are *Championship, Team, State, City* and *TwitterAccount*. Some examples of the concept properties that relate these concepts with each other are:

- `<http://khaos.uma.es/big data/kb/ont/play>` Property to connect two teams playing a match.
- `<http://khaos.uma.es/big data/kb/ont/cityIn>` Property to indicate relationships between cities and states.
- `<http://khaos.uma.es/big data/kb/ont/locatedIn>` Property to define relations between universities and cities.
- `<http://khaos.uma.es/big data/kb/ont/isAccountOf>` Property to indicate the twitter accounts of the teams.
- `<http://khaos.uma.es/big data/kb/ont/name>` Property to define the names for the entities.

B. Tweet Discovery

The first step of the proposed methodology is a well-known problem, *how to collect the tweets for their analysis*. The extraction of the tweets is done by means of Twitter API [11]. More specifically, in this case we have used Python's API [12]. The discovery component searches for certain keywords on Twitter. The keywords used are intended to obtain the highest number of tweets in order to have as wide as possible coverage, even if this means having to filter out some of them. These keywords are based on *hashtags* of interest and usernames related to the analysis task, but new keywords are detected during the analysis.

In our case, we have chosen as keywords the name of the user accounts of the university teams and hashtags created by team supporters. The championship also has its own hashtag. Examples of an initial set of keywords are *Big12Conference, Big12, Big12Insider, Big12MBB, BaylorMBB, CycloneMBB, TexasMBB, KUHoops, TCUBasketball, OSUMBB, kstatesports, kstate gameday, OU_MBBall, WVUhoops, TechAthletics and TTRaidersSports* [13-21]. We have compiled a wealth of information on approximately 11.5 gigabytes of tweets using these keywords.

Hadoop technologies have been used to deal with the retrieval and analysis of these tweets. The approach uses MapReduce, storing the tweets and the analysis results in HDFS (Hadoop Distributed File System). The following section details the analytical algorithms developed.

C. Ontology-based Tweet Analysis with MapReduce

The algorithms created to analyse the tweets make use of Natural Language Processing and Semantic Web techniques. These techniques take part of the MapReduce model that divides the process into three main functions: the Map function, the Combiner function and the Reduce function.

(1) The Map function retrieves a tweet and analyses its entities' sentiment value. Entities discovered by GATE are annotated as an RDF document. In the test case, GATE searches all the teams listed in the tweet (can be more than one), together with the sentiments associated with them in any individual tweet. SentiStrength calculates the sentiment of an entity based on its context. Hence, a tweet can deliver several sentiment values for each different entity.

The output of the Map function, which is the input of the Combiner function, is the tuple *(key, value)* whose contents are:

- *key*. It is a string formed by the joining of the tweet identification number concatenated with the date and time of the publication of the tweet.
- *value*. It is an RDF triple that is calculated in this Map function.

(2) The Combiner aims to group triples by team. The team associated with each entity is calculated during the execution of the Map function. This function takes care to gather together all positive and negative sentiment values from a particular entity (team in the use case) on a particular date and at a specific time (the date is given as day, month, year, hours, minutes and seconds). The time mark-up of the sentiment values is useful when analysing the results and their correlation with real events. In this case, they could be a score, the elimination of a team, injuries of a player, etc.

The output of the Combiner function is the entry for Reducer function. Therefore, outputs tuples *(key, value)* are:

- *key*. It is the joining of the entity name with the date and time at which the tweet was posted. In the use case, the entity name corresponds to the team name.
- *value*. It is the sum of the positive and negative sentiments calculated from the tweet.

(3) The Reducer collects the values of positive and negative sentiments calculated in the Combiner function and calculates the average of these values for all the tweets using entities (teams in the use case) and dates to order them. Thus, the output of this function is set of tuples *(key, value)* that are constructed as follows:

- *key*. It is the same key as the Combiner function.
- *value*. It is calculated by concatenating the number of tweets, positive sentiment value mean and negative sentiment value mean.

VI. Experimental Results

In this section, we examine the data obtained in the analysis phase. The objective of this analysis is to determine whether the results obtained during the Big 12 Men's Basketball Championship can influence the sentiments of tweets. Therefore, we look to analyse the feelings of the tweets before, during and after matches to check how they change over the course of the match.

The basketball championship began on Wednesday, March 12th, 2014 and ended on Saturday, March 15th, 2014. We started with the capture of tweets to analyse their feelings a few days before the start of the championship and completed the collection one day after it. This was done with the idea of assessing how the championship was affected by the tweets about the teams playing the tournament.

A. Global trends in the championship

In this subsection, we analyse the trends in the number of tweets and their sentiment values throughout the whole championship.

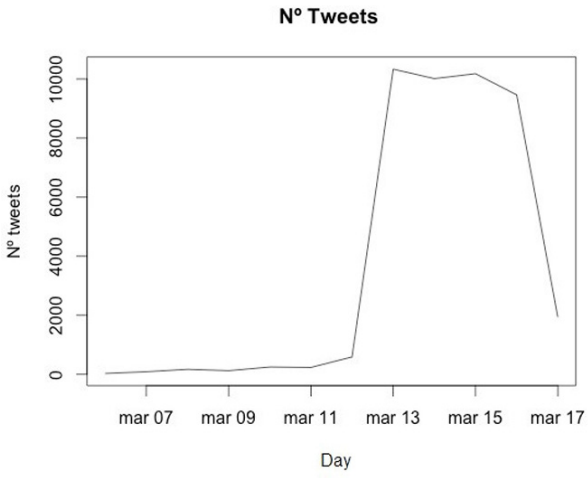


Fig. 2. Total number of tweets mentioning university teams or tournaments

Fig. *2* shows that prior to the start of the championship the number of tweets which mention the university teams or their tournaments were very low. Nevertheless, we note a significant increase in the number of tweets during the championship. Finally, this number falls as soon as the championship ends to a similar number as the initial phase of the championship.

Similarly, the feelings of the tweets, overall, increase their values when the championship starts. This is expected since the mood swings are intensified whenever there is a sporting event. These trends can be observed in Fig. 3 (positive) and Fig. *4* (negative).

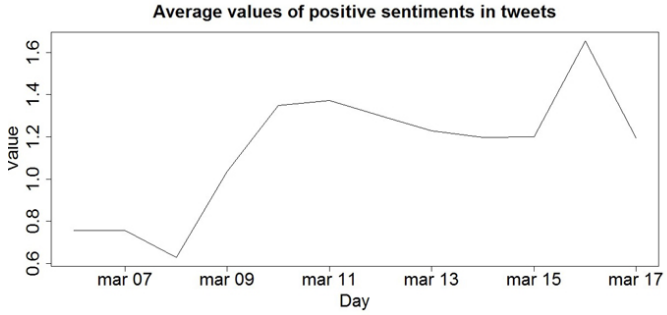


Fig. 3. Positive trends in tweets during the championship

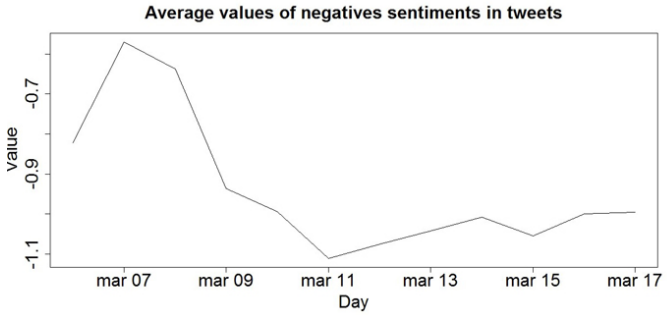


Fig. 4. Negative trends in tweets during the championship

B. Final match

Having viewed the data of the entire championship, we focus on one match specifically so as to compare the changing feelings for the two teams, so we choose the most important match of the championship which is the final.

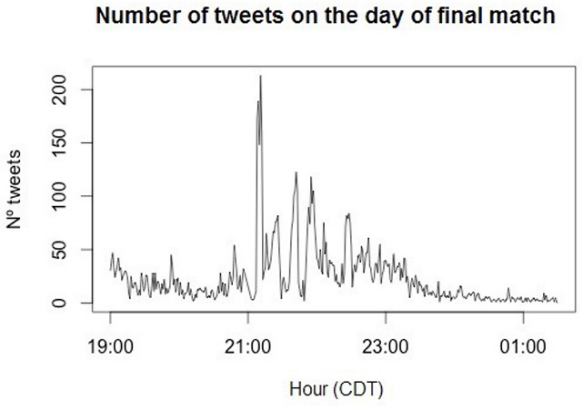


Fig. 5. Evolution in the number tweets in the final match

The final match was played by Baylor University against Iowa State University. Iowa State was the winner. The match began at 20:10 and

ended more or less at 22:10. As shown in Fig. *5*, at the time of the start the game, the number of tweets commenting on something about the championship begins to increase. In the time period between 22:00-22:30, the number of tweets starts descending. The maximum number of tweets is generated in the interval from 21:15 to 21:20. This interval coincides with half-time.

For a better understanding of the trend in the tweets, it should be noted that Baylor University was winning the game until the final minutes. However, Iowa State University, in the final minutes, took the lead and finally won the match.

In Fig. *6* we can observe the trend of positive tweets about Baylor University. A notable aspect is as the outcome becomes less favourable the sentiments in the tweets become less positive.

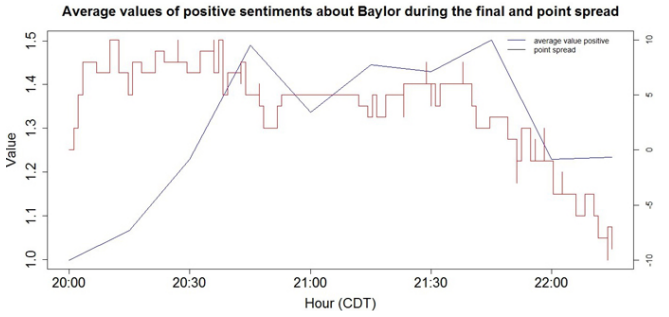


Fig. 6. Baylor University sentiment tweet's trend in the final match

However, we note that in the case of Iowa State University, the trend in tweet sentiments is different from Baylor, when the match is getting closer to the end, the sentiment is mostly positive, as we see in Fig. *7*. This is to be expected as Iowa won the match in the final minutes; therefore, their fans were happier and wrote positive tweets.

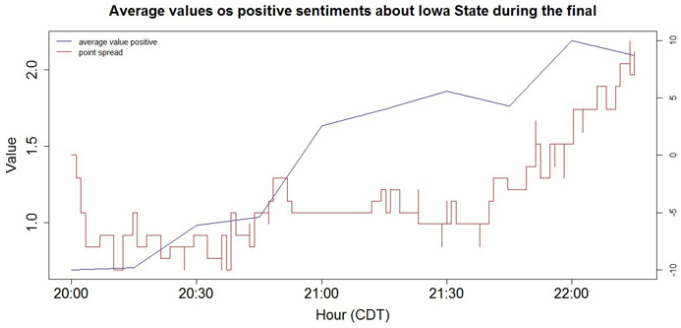


Fig. 7. Iowa State University sentiment tweet's trend in the final match

C. Regression Analysis

This section mathematically analyses the relationship between the number of tweets that are generated, and the positive or negative feelings that they show. This study focuses on the final match.

Linear regression is a mathematical method that models the relationship between pairs of variables. The Pearson product-moment correlation coefficient [24] measures the intensity of this possible relationship between variables. This rate applies when the relationship that may exist between the variables is linear and is calculated as the ratio between the covariance and the product of the standard deviations of both variables (Eq. 1).

$$r \quad \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Eq. 1. Linear regression.

The values can take a correlation coefficient range from -1 (perfect negative correlation) to 1 (perfect positive correlation).

$$-1 \leq r \leq 1$$

As we see in Tables I and II variables which have a better relationship with each other are the positive (+) and negative (-) sentiments. In Table III we note a relationship as a moderate positive sentiment between the number of tweets of Baylor and Iowa State during the final. In Tables IV and V, we can statistically see, as in the case of the Baylor correlation- though not as strong- a negative meaning so that when you increase the point spread against Baylor the number of tweets and sentiment values also decrease. In contrast, we note that in the case of the Iowa State University, when the point spread is higher the number of tweets and positive sentiment are also higher, so the correlation value is positive. However, for Baylor's negative sentiments as well as positive ones, the correlation coefficient is negative.

TABLE I
LINEAR REGRESSION ANALYSIS OF BAYLOR TWEETS. POSITIVE SENTIMENT IS DENOTED WITH (+) AND NEGATIVE SENTIMENT (-)

Baylor	Nº Tweets Baylor	Average (+) Baylor	Average (-) Baylor
Nº Tweets Baylor	N/A	-0.04865	-0.01833
Average (+) Baylor	-0.04865	N/A	0.53622
Average (-) Baylor	-0.01833	0.5363	N/A

TABLE II
LINEAR REGRESSION ANALYSIS OF IOWA TWEETS. POSITIVE SENTIMENT IS DENOTED WITH (+) AND NEGATIVE SENTIMENT (-)

Iowa State	Nº Tweets Iowa State	Average (+) Iowa State	Average (-) Iowa State
Nº Tweets Iowa State	N/A	0.40802	0.28161
Average (+) Iowa State	0.40802	N/A	0.74019
Average (-) Iowa State	0.28161	0.74019	N/A

TABLE III
LINEAR REGRESSION OF BAYLOR IOWA STATE TWEETS. POSITIVE SENTIMENT IS DENOTED WITH (+) AND NEGATIVE SENTIMENT (-)

Baylor/Iowa State	Nº Tweets Iowa State	Average (+) Iowa State	Average (-) Iowa State
Nº Tweets Baylor	0.56581	0.31094	0.225542
Average (+) Baylor	0.11722	0.27056	0.22139
Average (-) Baylor	0.00112	0.15629	0.143035

TABLE IV
LINEAR REGRESSION BAYLOR AND POINT SPREAD

Baylor	Nº Tweets Baylor	Average (+) Baylor	Average (-) Baylor
Point spread	-0.12698	-0.22582	-0.06735

TABLE V
LINEAR REGRESSION IOWA STATE AND POINT SPREAD

Iowa State	Nº Tweets Iowa State	Average (+) Iowa State	Average (-) Iowa State
Point spread	0.42327	0.36465	-0.22899

VII. Discussion

In this paper, a methodology that combines an ontology-based NLP process with a sentiment analysis to produce an accurate analysis of sentiment has been proposed and explained. The philosophy of the MapReduce approach fits perfectly in this type of work, because it allows us to distribute the workload of calculating the sentiment of a set of tweets into a computing cluster. As the calculation of feeling in a tweet is independent of other tweets, distributing the tweets can be done seamlessly without dependencies.

The fine grain of the sentiment calculation enables an analysis at different levels. In this paper, we have shown a use case related to a popular sporting event to show the feasibility of the proposal and the possible analysis that can be performed using the calculated sentiment. However, generating the information extracted and calculating sentiment as ontology instances enables them to be connected with other tools or scenarios:

1. The use of semantics in the analysis process enables the use of context in the discovery of entities in the tweets and the differentiation of the sentiment depending on the entities we are analysing.
2. The use of RDF triples facilitates the publication of the sentiment analysis as SPARQL endpoints, which in turn enables them to be linked with other Linked Data repositories.
3. The fine grain analysis allows a deeper analysis and even the building of data warehouses to operate with different filtering or aggregation functions.

The use case selected has helped test the approach in a real scenario where real events are used to contrast the predicted results with expected sentiment values. In this case, the feelings associated with the tweets correlate to the changes in the sporting event. Thus, the value of this type of analysis has been demonstrated for any context, where one-off events can alter the feelings of social communities and maintain these feelings sometime after the event that produced the changes in sentiment.

This work has focused on the analysis of the sentiment for individual tweets. However, the social interactions between Twitter users could be also of interest in this analysis [24]. Thus, future work will include the consideration of these social interactions to fine tune the analysis results.

VIII. Conclusions

This paper has presented a novel approach for the sentiment analysis of tweets using semantics. This work can be applied in other scenarios with small texts, by defining the analysis context with a domain ontology.

The use of sporting events for testing purposes has been shown to be valid, and so it can be used as a base line for the development of a benchmark for empirical testing of fine grain sentiment analysis tools.

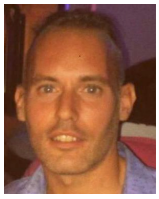
The parallelisation of the problem using MapReduce has shown a good behaviour for the analysis task. However, we are currently looking at ways to improve it. In future work we will be developing approaches, such as Apache Spark or Apache Storm, which are able to deal with real time analysis. These analyses with the support of a parallel platform will lead us to savings in terms of computational effort.

Acknowledgment

This work was supported in part by Grants TIN2014-58304-R (Ministerio de Ciencia e Innovación) and P11-TIC-7529 and P12-TIC-1519 (Plan Andaluz de Investigación, Desarrollo e Innovación). Cristobal Barba is supported by Grant BES-2015-072209 from the Spanish Government.

References

[1] Twitter Statistics, Last Access 28th January 2016. https://about.twitter.com/es/companyB.
[2] Big 12 Men's Basketball 2014 phillips 66 Big 12 Men's Basketball Championship. http://www.big12sports.com/.
[3] Apache Hadoop, 2016 Apache Hadoop http://hadoop.apache.org.
[4] Framework MapReduce, 2008 MapReduce Tutorial
[5] General Architecture for Text Engineering. GATE 2016. https://gate.ac.uk/
[6] SentiStrength . 2013 SentiStrength { sentiment strength detection in short texts sentiment analysis, opinion mining. http://sentistrength.wlv.ac.uk
[7] Luciano Barbosa,Junlan Feng. 2010. Robust Sentiment Detection on Twitter from Biased and Noisy Data. Coling 2010:Poster Volume, Beijing.
[8] S. Agarwal, S. Godbole, D. Punjani y S. Roy. 2007. How much noise is too much:A study in automatic text classication. ICDM, pages 3-12.
[9] Akshi Kumar, Teeja Mary Sebastia. 2012. Sentiment Analysis on Twitter. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012.
[10] Twitter4j. 2016 Api Twitter4j. http://twitter4j.org/en/index.html
[11] Twiter Api, 2016 https://dev.twitter.com/overview/api
[12] Python Api, 2016 https://docs.python.org/2/c-api/
[13] Baylor University Baylor University Texas http://www.baylor.edu.
[14] Iowa State University Iowa State University..
[15] Kansas University Kansas University http://www.ku.edu.
[16] Kansas State University Kansas State University http://www.k-state.edu.
[17] Oklahoma University Oklahoma University https://www.ou.edu.
[18] Oklahoma State University Oklahoma State University http://go.okstate.edu.
[19] Texas Christian University Texas Christian University http://www.tcu.edu.
[20] Texas Tech University Texas Tech University http://www.ttu.edu.
[21] West Virginia University West Virginia University http://www.wvu.edu.
[22] The Pearson product-moment correlation coefficient, 2016
[23] Staab, S., & Studer, R. (2009). Handbook on ontologies. International Handbooks on Information Systems. Springer.
[24] Teófilo Redondo. The Digital Economy: Social Interaction Technologies – an Overview. *International Journal of Interactive Multimedia and Artificial Intelligence*. 3(2): 17-25.
[25] Monchón.F, SanJuan (2014). O. *A First Approach to the Implicit Measurement of Happiness in Latin America Through the Use of Social Networks*. International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 2, Nº 5



Cristóbal Barba-González is a PhD student at the University of Málaga. His research focuses on optimization algorithms (metaheuristics), semantics and analysis of big data.

José Manuel García-Nieto is a PhD and Post-Doctoral Research Assistant at the University of Málaga. His research focuses on optimization algorithms (metaheuristics), semantics, data mining and Big Data.

Ismael Navas-Delgado is PhD and Assistant Professor at the University of Málaga, Spain. His research focuses on the use of Semantics and Big Data technologies in Life Sciences.

José F. Aldana-Montes is Full Professor at the University of Málaga, Spain. He is the main researcher of the Khaos Research Group. His research focuses on management, integration and analysis of data.