Prediction of Event Attendance for FIFA 14 World cup

David Perikala
University of Passau
nagilladavid@hotmail.com

1 INTRODUCTION

Study of how a large community or a population behaves in a major event is a primary question, that in the past was extremely difficult to approach on a large scale [7]. With the growth of social networks such as Facebook, Instagram, Foursquare and Twitter, it became possible to analyse the real-time behaviour of large groups of people attending popular events [7][2].

1.1 Problem Statement

This paper presents a study of predicting the attendance of Twitter users at the FIFA 2014 World cup event based on geotagged and non-geotagged posts. We propose a machine learning approach for analysing social media posts of users discussing the event to predict their attendance.

1.2 Proposed Solution

In this paper, we present the study of twitter data by deriving a machine learning model which predicts the attendance of the user. The data consists of tweets collected during the 64 matches of the World Cup from June 12 to July 13, 2014 [4]. Our approach to the above stated problem is to consider tweets with and without the location tags for which we state a ground truth for predicting if the user has attended or not attended the match. We intend to implement a supervised learning approach for our study. Based on the tweets we retrieve, only two percent of that data have geo-locations [2]. We have established a set of ground truth for our study so as to narrow down our scope. These rules can be described as following:

- From the data we gather, we consider only the date and not the time of tweet, hence the exact time of tweet is eliminated.
- We focus only on the tweets which are in English language, and remove the other language tweets.
- For our location assumption, we consider tweets based on the country and not as exact point location of match for various reasons such as: all matches are not at one location, people are constantly traveling following their team, usually people stay far away from venue of the match depending on city capacity even though taking radius from venue would have worked; but we avoided it since most of the cities for the World Cup were near border cities such as Fortaleza, Natal, Recife , Salvador, Porto Alegre so taking a radius might change country location. Hence, we considered that if the tweet location is traced from Brazil, the user is attending, else not attending.
- For tweets without geo location we look for tweet text and search root word 'attend' or 'watch', if these words are found within the text, we consider as attending; remaining tweets are scrapped off to make dataset balanced.

We generated labeled data from tweets on the following basis:

(A) For tweets with geo-location
If a user is tweeting from location 'Brazil' and during the world cup duration defined above, we assign the label as 'Attending', else with any other locations we assign the label as 'Not attending'.

(B) For tweets without geo-location
Here, we consider the tweet text, if the text contains root words like 'watch' or 'attend', we assign the label as 'Attending', else assign the label as 'Not attending'

In the following sections, the proposed solution is explained in detail.

1.3 Proposed Evaluation

We propose to use confusion matrix to describe the performance of our classifiers, for which the number of true positives, true negatives, false positives and false negatives are calculated. The standard precision, recall, and F1 scores are then computed accordingly. We also evaluate our assumption of assigning labels by randomly and manually selecting 50 tweets and then deriving a confusion matrix based on our readings.

2 METHODOLOGY

In this paper, we present our study and achieved results based on the use of methodology designed for the collection and analysis of tweets. The main goal of this work was to monitor the attendance of Twitter users during the FIFA World Cup 2014 matches. The data gathered includes all tweets collected during the 64 matches of the World Cup from June 12 to July 13, 2014. For each match, we consider both the geotagged tweets throughout the event. Original results were obtained in terms of number of people who attended the event. We have the whole process composed of four main steps: data acquisition, data pre-processing, data mining and results.

2.1 Data acquisition

Data acquisition has been carried out by referring crawled data, which we found during our survey in the data gathering phase. We downloaded a CSV file containing Tweet IDs that were crawled during the FIFA 14 World Cup (reference link mentioned). This CSV file contains about 30 million Tweets (80 million including retweets which are omitted) collected between 6th June and 14th July 2014. These Tweets were filtered by some official World Cup hashtags such as #WorldCup2014, #Brazil2014, #FIFA2014, as well as some country codes including #GER for Germany, #ARG for Argentina #BRA for Brazil, #GHA for Ghana and so on [4].

2.2 Data preprocessing

Pre-processing has been performed to clean, select and transform data to make it suitable for analysis. First, we cleaned collected data by removing all the invalid tweets. Using Tweepy, the python

library for accessing the Twitter API, we validated the tweet IDs and considered only the valid ones for further processing. The validation step was initially executed on the entire CSV file containing the twitter IDs, although due to time limitation and greater computational power needed on such huge amount of data, processing the entire file at once was not feasible. Due to this huge count of 30 million tweets, the validation step was not completed even after 72 hours of processing and hence we adopted a different approach, that is to split the data into smaller files. After the split, we had 1080 files each with approximately 30000 tweet IDs. We limited our research to focus only on 200 of these files as processing such huge data needs lot of computational power and time. Each file was then processed on parallel systems for identifying the valid tweets. Post the execution of validation step on these files, we had gathered around 18-22k valid tweets per file (average of 20k) giving us nearly 4,000,000 valid tweets. The gathered data was then used to derive a set of tweets, where each tweet is described by following properties: tweet ID, location co-ordinates (latitude and longitude), date and text. Valid tweet IDs were filtered based on event dates (12 June to 13th July 2014). The tweet location was identified using the 'GeoPy' which is a geocoding library for python. In our implementation, we adopted the 'Nominatim' geocoding web service from GeoPy. We adopted this web service because it generates the address by reverse geocoding, that is generate an address (location name) from a latitude and longitude, which was our exact requirement.

2.3 Data Mining

A data mining task was performed for the prediction of event attendance and also in implementing machine learning classifiers. One of the most common tasks in machine learning is text classification, which is simply teaching machine how to read and interpret a text and predict what kind of text it is. In our study, we have supervised learning approach, where we fed the learning algorithm with a training data X and its already known output Y (because it's a training data pair (X, Y)), which is, for a given text X, Y is its classification, the algorithm will learn it. The data partition (training data and testing data) is done using the 'train_test_split' method from Python's Scikit Learn, where the split is 20% for test and 80% for training. Before executing the classifier algorithms, we prepare the data by representing labels (attending and not attending) and tweets through term frequency and inverse document frequency (TF-IDF). TF-IDF is a way to convert textual data to numeric form and is short for Term Frequency-Inverse Document Frequency. Using the Scikit Learn again, we use the 'TfidfVectorizer' method, this method will do the TF-IDF on the data, and then vectorize this data, in other words, transform the whole thing into an array of inverse frequencies. So, extracting these features we get the training and test data, that we feed the algorithm along with the Y. Now we have, X which is the set of features (vector with the TF-IDF of each data point) and Y is the output, which is, the labels/class (e.g: attending or not-attending). We further use the label encoder class and its fit transform method, the idea is that, we only want numeric and continuous values in the dataset. It is quite simple to convert using encoder in python. Encoder will convert the text in the dataset into numeric value (0 and 1). With the data prepared, selected and features extracted, we feed the algorithm with this data. In our experiment, we implement multiple algorithms to understand the learning of different classifiers and their behavior when input same data. The classifiers include Logistic Regression, Naive Bayes, Decision tree and SVM (support vector machine). These algorithms were chosen using the best of our knowledge by considering the size, quality, and nature of the data along with training time required for the classifier.

Logistic regression (Predictive Learning Model):. We implement logistic regression because it performs binary classification, so the label outputs are binary which is similar to our case that is the two classes - attending and not attending. It is known to be fast and simple for execution.[3]

Naive Bayes Classifier (Generative Learning Model): It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.[1]

Decision tree: Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The result is a tree with decision nodes and leaf nodes. Each node represents a single input variable (x) and is split point on that variable. The leaf nodes of the tree contain an output variable (y) which is used to make a prediction. Predictions are made by walking the splits of the tree until arriving at a leaf node and output the class value at that leaf node.[6]

SVM: In this algorithm, we plot each data item (tweet) as a point in n-dimensional space (where n is number of features) with the value of each feature being the value of a coordinate (Support Vectors). A hyperplane is a line that splits the input variable space. We apply SVM learning algorithm because it finds the coefficients that results in the best separation of the classes by the hyperplane. [5]

3 RESULTS

The study aims at predicting the attendance of Twitter users during the FIFA World Cup 2014 matches. In general, we extracted the information and establish ground rules to identify if the user tweeting about the match has attended the match or not. In the following, we present an extract of our results:

(A) Manually calculating the confusion matrix.
To understand how well have we trained our classifiers, we randomly selected 50 tweets. The idea was to create a confusion matrix by manually studying how well the classifiers have performed.

N = 50		Predicted	
		Not Attending	Attending
Actual	Not Attending	16	4
	Attending	5	25

Table 1: Results by manually evaluating the prediction of classifiers

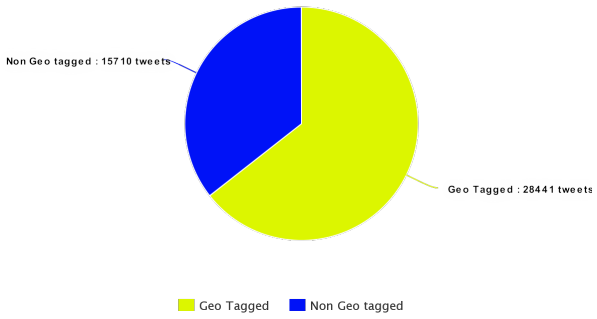

Fig 1: Distribution of tweets

Here is an example of misclassified data:

(i) **Tweet "Neymar stopped watching Brazil vs Germany 7th goal played poker instead world cup 2014" was labeled as Attending**

After reading the above tweet, we can state that it was wrongly classified as 'attending' as the person was definitely not watching the Brazil vs Germany match. (The above tweet is an outcome of preprocessed data hence the hash-tags and the stop words are removed making it a little unreadable).

From the 28441 geo-tagged tweets, 27466 were identified to be outside the geographic area of Brazil, such tweets are labeled as 'Not Attending'. The remaining 975 tweets were labeled as 'Attending' because they are identified within the geographic area Brazil.

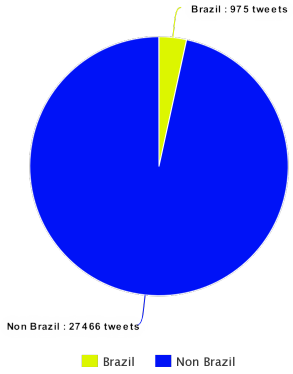

Fig 2: Distribution of geo-tagged tweets

Accuracy	82%
Precision	86.2%
Recall	83.33%
F1 Score	84.7%

Table 2: Evaluation metrics

(B) Classifier results

We input labeled file to our classifier for training purpose. We had approximately 44k data out of which 28441 were labeled on the basis of geo tags, i.e. if user is in Brazil then labeled as "Attending"' and if not in Brazil "Not attending". The remaining 15710 tweets with non geotag locations were labeled as 'Attending' based on the words "watch" or "attend" present in tweet text.

Since most of the tweets labeled as attending were generated on the basis of the words 'watch' and 'attend', it makes the learning of classifier easy by identifying the presence of the defined words; it is because of this that the classifiers generate results with higher accuracy. In order to study the learning of classifier, we scrap the words 'attend' and 'watch' from the file which would be input to the classifiers, this file contains the tweet text. We re-train our classifiers and generate results for a comparative study before and after the removal of words. (Kindly refer report 3 for comparison of the results before and after scrapping the words.)

Classifier	Accuracy	F-Measure	Precision	Recall
Naive Bayes	72.99	70.68	75.922	66.12
SVC	76.26	74.07	75.59	72.60
Logistic Regression	75.92	73.54	71.84	71.84
SGD	69.78	67.48	67.48	67.22
Random Forest	76.04	73.76	75.37	72.21

Table 3: Comparative study of the classifiers based on accuracy, F-measure, Precision and Recall

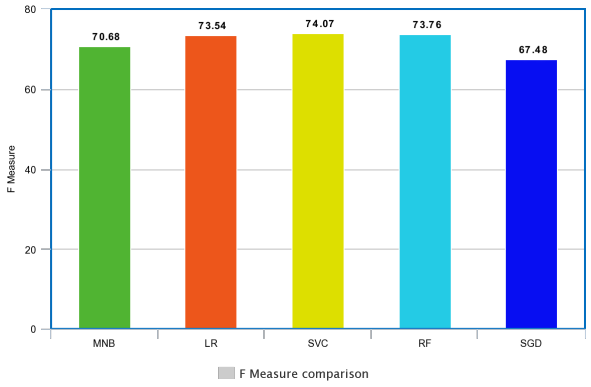

Fig 3: F-measure comparison between classifiers

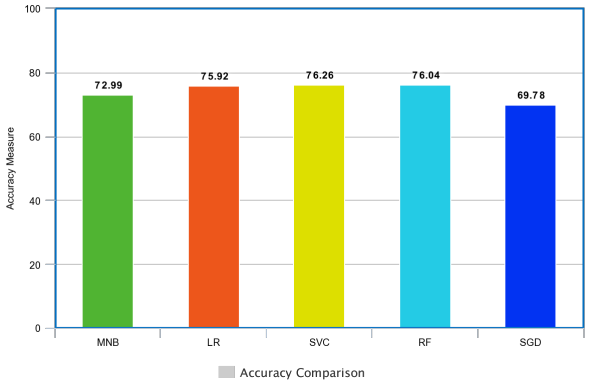

Fig 4: Accuracy comparison between classifiers

From the above figure, it shows that SVC performed the best for our dataset and SGD performed the worst amongst the 5 classifiers implemented.

4 CONCLUSION & FUTURE WORK

In this study, we explored the aspect of event attendance of a sporting event -FIFA 2014 World Cup- using data from Twitter. As further analysis, we would classify tweets with images and extract EXIF details of these images in order to get more tweets from Brazil location as this would increase the count of tweets with geo locations as parameter for labeling. For the Language filter, additionally we include Spanish tweets as they were the second most tweet in any language[4]. This would help in making more accurate labeling.

In future, we can build on this research. As sporting events occur frequently, there exists lots of opportunities for exploitation. With the increase in the usage of social media and technically advanced smart phones, much more research can be done based on machine learning which helps us to identify many new features such as extracting the exact location, movements of the user, frequency of the tweets and also to predict the favorites of a game based on sentiment analysis of a particular match.

REFERENCES

[1] Naive Bayes. 2018. URL address: https://en.wikipedia.org/wiki/Naive_Bayes_classifier/.

[2] Vinicius Monteiro de Lira, Craig Macdonald, Iadh Ounis, Raffaele Perego, Chiara Renso, and Valeria Cesario Times. 2017. Exploring Social Media for Event Attendance. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, 447–450.

[3] Logistic Regression. 2018. URL address: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-logistic-regression/.

[4] Data Source. 2014. URL address: http://blog.aylien.com/text-analytics-meets-2014-world-cup-tweets-part-1/.

[5] SVM. 2018. URL address: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-support-vector-machine/.

[6] Decision Tree. 2018. URL address: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-boosted-decision-tree/.

[7] Zhijun Yin, Liangliang Cao, Jiawei Han, Jiebo Luo, and Thomas Huang. 2011. Diversified trajectory pattern ranking in geo-tagged social media. In *Proceedings of the 2011 SIAM International Conference on Data Mining*. SIAM, 980–991.