Extracting Semantic Entities and Events from Sports Tweets

Smitashree Choudhury[1], John G. Breslin[2]

[1] DERI, National University of Ireland, Galway, Ireland
[2] School of Engineering and Informatics, National University of Ireland, Galway, Ireland
smitashree.choudhury@deri.org, john.breslin@nuigalway.ie

Abstract. Large volumes of user-generated content on practically every major issue and event are being created on the microblogging site Twitter. This content can be combined and processed to detect events, entities and popular moods to feed various knowledge-intensive practical applications. On the downside, these content items are very noisy and highly informal, making it difficult to extract sense out of the stream. In this paper, we exploit various approaches to detect the named entities and significant micro-events from users' tweets during a live sports event. Here we describe how combining linguistic features with background knowledge and the use of Twitter-specific features can achieve high, precise detection results (f-measure = 87%) in different datasets. A study was conducted on tweets from cricket matches in the ICC World Cup in order to augment the event-related non-textual media with collective intelligence.

1. Introduction

Microblogging sites such as Twitter[1], Tumblr[2] and Identi.ca[3] have become some of the preferred communications channels for online public discourse. All of these sites share common characteristics in terms of their real-time nature. Major events and issues are shared and communicated on Twitter before many other online and offline platforms. This paper is based on data obtained from Twitter because of its popularity and sheer data volume. The amount of content that Twitter now generates has crossed the one billion posts per week mark from around 200 million users, covering topics in politics, entertainment, technology and even natural disasters like earthquakes and tsunamis. Extracting useful information from this constant stream of uninterrupted but noisy content is not trivial.

[1] http;//www.twitter.com/
[2] http://www.tumblr.com/
[3] http://www.identi.ca/

The extraction of useful content such as entities, events and concepts needs to address many conventional IR-related issues as well as some Twitter-specific challenges. Nevertheless, the results can be useful in many real-world application contexts such as trend detection, content recommendation, real-time reporting, event detection, and user behavioural and sentiment analysis, to name a few. In the present study, we tried to detect named entities and interesting micro-events from user tweets created during a live sports event (a cricket match). The application of these results aims to augment sports-related multimedia content generated elsewhere on the Web.

Making sense of social media content is not trivial. There are many social media-specific challenges in capturing, filtering and processing this content. Some of the typical issues are as follows:

- Tweets are 140 characters in length, forcing users to use short forms to convey their message. Many routine words are shortened such as “pls” for “please”, “forgt” for “forgot”, etc. We need a special dictionary to understand this constantly-evolving community-specific lingo.
- There is a lack of standard linguistic rules. Due to the lack of space, language rules are avoided when necessary, and as a result conventional information extraction techniques do not work as expected.
- The use of slang words, abbreviations and compound hashtags are community driven rather than based on any dictionary or knowledge base.

The goal and objective of this paper is to classify the tweets mentioning the named entities and interesting events occuring during a live game. Despite knowing that the content generated during an event includes discussions and opinions about the event, detecting the discussed entities and interesting sub-events is challenging. As an example, consider a tweet “O'Brien goes ARGH!!!” which actually means that a player called (surname) O'Brien got out. Manual observation says that this tweet contains one named entity (the player's name) and one interesting event (getting out), but text processing applications fail to detect them due to the lack of context rules. We propose various approaches including linguistic analysis, statistical measures and domain knowledge to get the best possible result. For instance, instead of simple term frequency measures, we represent each player and possible interesting events with features drawn from multiple sources and further strengthen their classification score with various contextual factors and user activity frequency (tweet volume).

Our contribution includes:

- Detecting named entities based on various feature sets derived from tweets and with the help of background knowledge such as event websites and Wikipedia.
- Developing a generic framework to detect interesting events which can be easily transferred to other sports events.

Figure 1 shows a visual illustration of the steps followed in this work.

The rest of the paper is organised as follows: section 2 presents our methodology and approaches to address the issues of feature selection and classification; section 3 describes the evaluation and results of the study. Related work is discussed in section 4, followed by conclusions in section 5.

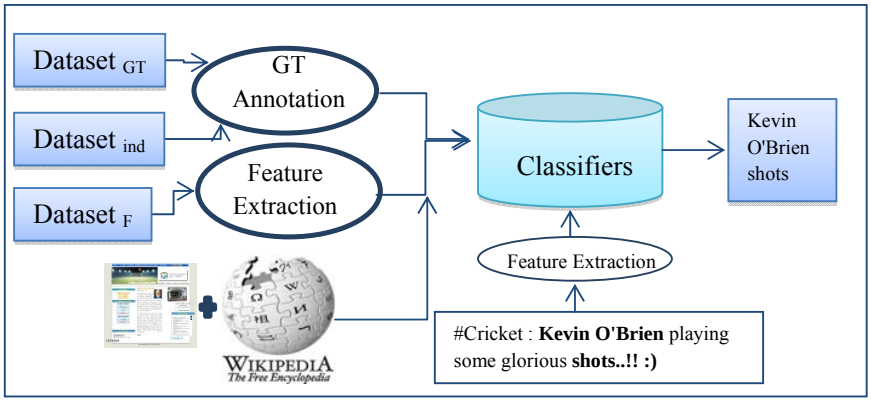

Fig.1. Overview of various steps followed.

2. Methodology

Our goal is to build classifiers which can correctly detect the players' named entities and the interesting micro-events within a sports event. We started by crawling tweets during the time of the cricket matches using the Twitter API. Since we can crawl tweets with keywords, we collected some related keywords and various hashtags (ICC cricket world cup, #cwc2011, cwc11, cricket, etc.) as a seed query list. Despite our filtered and focused crawling, many users use the popular hashtags and keywords to spam the stream to get attention. Including these tweets due to the mere presence of hashtags or keywords may bias the analysis, so a further round of de-noising is performed following a few simple heuristics as described below:

1. Messages with only hashtags.
2. Similar content, different user names and with the same timestamp are considered to be a case of multiple accounts.
3. Same account, identical content are considered to be duplicate tweets.
4. Same account, same content at multiple times are considered as spam tweets.

Using the above heuristics, we were able to identify and remove 1923 tweets from the dataset of 20,000 tweets. Our goal is not to eliminate all noise but to reduce it as much possible in order to get a proportionally higher percentage of relevant tweets.

The next step is to divide the datasets into two parts ($D_{Feature}$ and $D_{GroundTruth}$). $D_{GroundTruth}$ is manually annotated and $D_{Feature}$ is used for feature extraction. Each event and entity is considered as a target class and is represented with a feature vector. Details of the feature vector are described in sections 2.3 and 2.4.1.

Once the players are represented with the feature vector, the next step is to classify the tweets to say whether it contains any mention of a player or not. If the classification is positive, then matching is performed based on the player's full name. Each player is considered as a target class. Let $P = \{p_1, p_2, \ldots p_n\}$ be a set of players and let $FV(p_i)$ be a set of features used to represent the player. Let $M = \{m_1, m_2, \ldots m_n\}$ be a set of tweets belonging to a single game. We then train the classifier:

$$f(pi, mi) = \begin{cases} 1 = if\ mi\ makes\ a\ reference\ to\ a\ player\ pi \\ 0 = if\ mi\ makes\ no\ reference\ to\ a\ player\ pi \end{cases}$$

where p_i is the player's feature and m_i is the input tweet. Similar classification is performed for the micro-event detection task.

2.1 Dataset

We collected three datasets for training, testing and feature selection. Dataset (D_F) is a collection of 20,000 messages collected during the first round matches of the ICC World Cup. Dataset D_{GT} is a subset of D_F and consists of 2000 tweets. Dataset $D_{independent}$ (D_{ind}) (independent of training) is a set of 1500 messages from one game played between Ireland and England. Dataset D_{GT} and D_{ind} are manually annotated with a label of the player's name for any player entities and with "yes", "no" or "others" for the presence or absence of interesting events. Three students with a knowledge of the game were asked to annotate D_{GT} and D_{ind}. To increase the quality, we gave them information regarding the matches they were looking at and also regarding the team players. To maintain the quality of annotations, we considered that two out of three annotators had to agree for a label. The results showed that all three agreed on labels in 86% of cases while agreement between two occurred 94% of the time.

2.2 Background Knowledge

Since the main event (a game between two teams) is a pre-scheduled event, we obtained the background knowledge - in terms of the team names, venue, date, starting time, duration, and player details (names) - from the game website. We also collected various concepts common to cricket games from Wikipedia as a list of context features. The list consists of domain terms such as "crease", "field", "wicket", "boundary", "six", "four", etc. All of this background information was collected manually.

2.3 Feature Selection for Entity Detection

We developed a player classifier which captures a few general characteristics and language patterns from the tweets. Each feature is given a binary score of 1, 0.

2.3.1 Terms Related to a Player: The vector consists of name-related features. These are: full name, first name only, last name only, initials, etc. One more feature which we considered to be useful was the nickname of the player. However, since correlating nicknames to player names proved difficult, we could not include that feature. Table 1 below shows a few examples of the feature subset.

Table 1: Features related to a player.

Player	**Name-Related Feature**
Kevin Peterson	<Kevin Peterson, Peterson, KP>
Sachin Tendulkar	<Sachin Tendulkar, Sachin, Tendulkar, SRT>

2.3.2 Terms Related to the Game: While studying the tweets, we realised that a player's name alone and its variations will lead to low precision as there may be many irrelevant discussions mentioning the player's name. In order to increase the quality and precision, we added a context feature where the game-related key terms appear within a window of four words. These key terms are manually prepared, which has been discussed in the background knowledge section. Examples of such occurrences are given below in Table 2. If we find these rules existing in the message, the feature score becomes 1.

Table 2: Tweets with the context feature.

#Cricket : **Kevin O'Brien** playing some glorious **shots..!! :)**
Captain Afridi goes this time**, wicket** for **Jacob Oram.**
First **SIX** of the tournament for **Afridi!!!** #cwc2011

As tweets are highly informal, capitalisation is infrequent, but when it does occur we count it as a feature and score accordingly. Many players are now addressed and mentioned via their Twitter account, so the presence of a player's username (@<player>) or hashtag (#<player>) are also counted as Twitter-specific features. Finally, a player's feature vector looks like:

$$FV(p_i) = \langle full\_name, first\_name\_only, last\_name\_only, initials, initial{+}lastname, context\_word, capitalisation, player\_mention, player\_hashtag \rangle$$

2.4 Micro-Event Detection

An event is defined as an arbitrary classification of a space/time region. We target events which are expected to occur during a certain time frame (i.e. the match duration), but location is not an issue here as we know the venue of the match and we are not interested in fine-grained locational information such as field positioning within the stadium. We made a few assumptions regarding an event's characteristics, namely that (1) they are significant for the results of the game, and (2) many users (the audience) will be reacting to these events via their tweets. The methodology options available for detecting game-related micro-events from tweets are: (1) statistical bursty feature detection; and (2) feature-based event classification. We combined both approaches to get the best possible result.

2.4.1 Event Feature Selection

Interesting events that arise during a game are not pre-scheduled, but there is the possibility that these events can occur at any moment of time during the game. We manually selected these events from the Wikipedia "Rules of Cricket" pages. There are two broad categories ("scoring runs" and "getting out") and 12 sub-categories of micro-events. Through our observation of tweets, we saw that most tweets referred to the "out" event by itself while not bothering too much with the specific "out" types such as "bowled", "LBW" or "run-out", though they are occasionally mentioned. Based on this, we restricted our classification task to three major possible events, i.e. "out", "scoring six", and "scoring four". Each event is represented with a feature vector which consists of keyword features related to the event.

Keyword Variations: An event is represented by various key terms related to the event. The logic of including such variations is that users use many subjective and short terms to express the same message - "gone", "departed", "sixer", "6", etc. – when caught up in the excitement of the game. These features are again extracted from the D_F dataset.

Linguistic Patterns: Like the player classifiers, the event classifier also includes contextual features and linguistic patterns to detect the events. The presence of such a pattern gets a score of 1 for the feature, otherwise 0. A few of the examples are shown below:

Table 3: Mentions of interesting events during a match.

#sixer from **#kevinobrien** for #ireland against #england #cricket
Kevin O'Brien OUT ! Ireland 317/7 (48.1 ov) #ENGvsIRE #cricket #wc11
Crap **O'Brien goes** ARGH!!!

2.4.2 Tweet Volume and Information Diffusion

We cannot say from a single tweet that an event has occurred. In order to make our detection reliable, we take crowd behaviour into account. Based on the assumption that interesting events will result in a greater number of independent user tweets, we computed two more features to add to the event feature vector: (1) the tweet volume; (2) the diffusion level. Tweet volume is the level of activity while the event is being mentioned, taken during a temporal interval tm_i where $i = \{1 \ldots n\}$ and the duration of each tm_i is two minutes (can be any duration depending the requirement). We used a two-minute interval for simplicity but it can be of any temporal size. If the number of messages is higher than a threshold of average plus 1 α, we mark the feature as 1, otherwise set it to 0.

The second feature is the level of information diffusion that takes place during the time interval tm_i. It is presumed that more and more users will be busy sharing and communicating the event through their own tweets rather than reading and forwarding others. This means that there will be less retweets (RTs) during the event interval compared to the non-event intervals. This assumption has been confirmed from our observations of the data that the immediate post-event interval has a lesser number tweets than the non-event intervals. The same assumption is also proved in the study [2]. The feature is marked the same way as the tweet volume feature.

3. Evaluation and Results

Our evaluation started with the dataset D_{GT} which is manually labelled both for players and interesting events. We first ran the players classifier and the results are shown in Figure 2. The objective of the evaluation is to judge the effectiveness of the proposed approaches to detect players' named entities and game-related micro-events against the manually-annotated datasets D_{GT} and D_{ind}. We also tested the weight of various features in classification (positive) and found that a combination of any name feature with the context feature (game-related term) is the best performing feature compared to any other combinations (Figure 5).

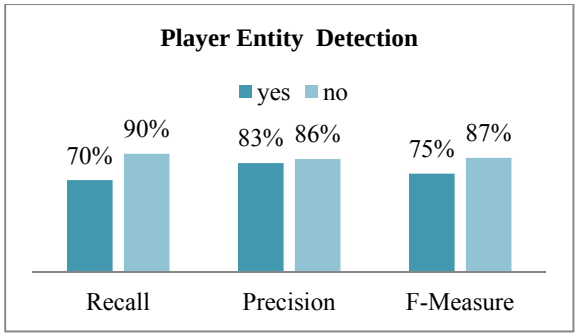

Fig. 2: Recall and precision of the player detection classifier.

Like the player classifier, we ran the same evaluation for micro-event detection but in two different stages: (1) classification with only linguistic features, and (2) classification with all features. With linguistic features only (Figure 3), recall is very low at 70% and precision is 74%. This may be due to the noise in tweets. Many event-related keywords are also used in normal conversations like "out", "over", etc.

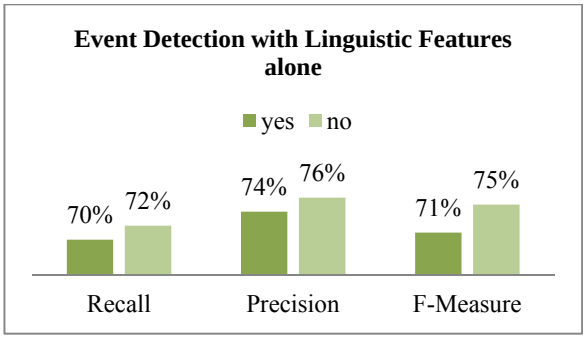

Fig. 3: Event detection performance with linguistic features only.

However, when we included the tweet volume and information diffusion level scores, both recall and precision further increased to 86% and 85% respectively, as shown in Figure 4.

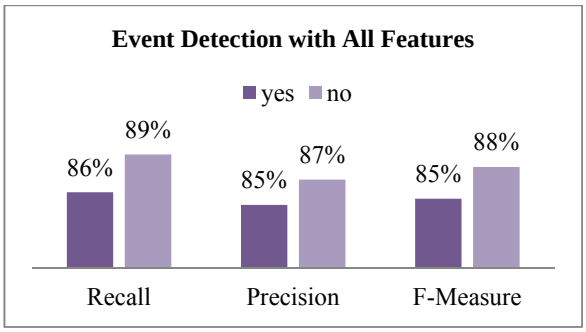

Fig. 4: Event detection performance with all features combined.

The results show that irrespective of any features, performance for the "no" labels is always better than for the "yes" labels. We assume this result may be due to the

greater number of negative samples available in the data compared to the positive samples.

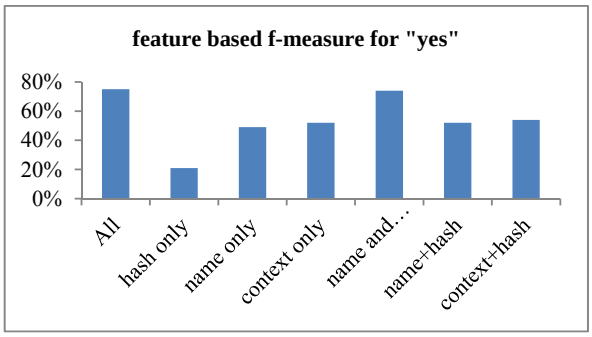

Fig. 5: Individual feature performance in player classification.

One question we were interested in answering was can the classifiers be used on other data which is independent of the training and the testing data? To explore this proposition, we ran the classifier on the independent dataset D_{ind} collected from a different game involving two different teams (England vs. Ireland). For this experiment, we tagged the content with part-of-speech tagging using the Stanford NLP tagger[4]; in the feature space, we replaced the player's name with a proper noun placeholder. A summary of the results for both players and event detection is shown Figure 6.

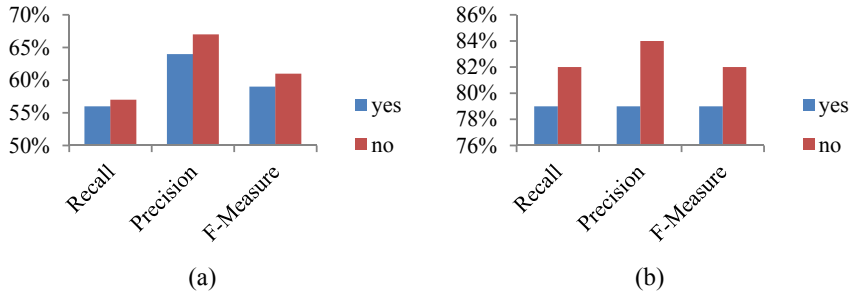

Fig. 6: (a) Player detection and (b) event detection in dataset D_{ind}.

As expected, the player classifier scored poorly compared to the event classifier, as the player classifier is heavily dependent on the players' names and their variations. Even if we replace the names with proper noun placeholders, many player mentions are only by first or last name, and other names could not be identified as proper nouns

[4] http://nlp.stanford.edu/links/statnlp.html

by the part-of-speech tagger. However, the event detection results are good, and the F-measure is above 80% as the features are more generic in nature.

4. Related Studies

Twitter is one of the most popular social media sites with hundreds of thousands of users sending millions of updates every day. It provides a novel and unique opportunity to explore and understand the world in real time. In recent years, many academic studies have been carried out to study issues such as tweet content structures, user influence, trend detection, user sentiment, the application of Semantic Web technologies in microblogging [1], etc. Many tools exist for analysing and visualizing Twitter data for different applications. For example, [3] analyses tweets related to various brands and products for marketing purposes. A news aggregator called "TwitterStand" is reported in [4] which captures breaking news based entirely on user tweets.

The present study addresses the research question of identifying named entities mentioned in microblog posts in order to make more sense of these messages. Therefore, the focus of our discussion in this section will be on various related studies concerning entity and event recognition in social media scenarios, especially in microblogs. Finin et. al [7] attempted to perform named entity annotation on tweets through crowdsourcing using Mechanical Turk and CrowdFlower. Similar research in [8] reported an approach to link conference tweets to conference-related sub-events, where micro-events are pre-defined as opposed to the sports domain where interesting events unfold as and when the event proceeds. Researchers in [2] built a classifier based on tweet features related to earthquakes and used a probabilistic model to detect earthquake events. Authors in [5] used content-based features to categorise tweets into news, events, opinions, etc. Tellez et al. [6] used a four-term expansion approach in order to improve the representation of tweets and as a consequence the performance of clustering company tweets. Their goal was to separate messages into two groups: relevant or not relevant to a company. We have adopted many lightweight techniques to identify named entities and micro-events during a sports event so that we can later use these results to address existing problems related to conceptual video annotation.

5. Conclusion

We presented approaches to identify named entities and micro-events from user tweets during a live sports game. We started with a filtered crawling process to collect tweets for cricket matches. We arranged three datasets (D_F, D_{GT}, D_{ind}); D_{GT} is a subset of D_F. D_{GT} and D_{ind} are manually annotated with player names and "yes" or "no" for players and events respectively, while D_F was used to extract the feature set. Classifiers built on these features were able to detect players and events with high precision. The generic features of our event detection classifier were applied to an independent dataset (D_{ind}) with positive results. Our future work includes transferring

the algorithm to other sports areas as well other domains such as entertainment, scientific talks and academic events.

Acknowledgments

This work was supported by Science Foundation Ireland under grant number SFI/08/CE/I1380 (Líon 2).

References

1. A. Passant, T. Hastrup, U. Bojars, J.G. Breslin, "Microblogging: A Semantic Web and Distributed Approach", The 4th Workshop on Scripting for the Semantic Web (SFSW 2008) at the 5th European Semantic Web Conference (ESWC '08), Tenerife, Spain, 2008.
2. T. Sakaki, M. Okazaki, Y. Matsuo. "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors", Proceedings of the 19th World Wide Web Conference (WWW2010), Raleigh, NC, USA, 2010.
3. B.J. Jansen, M. Zhang, K. Sobel, A. Chowdury, "Twitter Power: Tweets as Electronic Word of Mouth", Journal of the American Society for Information Science and Technology, 2009.
4. J. Sankaranarayanan, H. Samet, B. Teitler, M. Lieberman, J. Sperling. "Twitterstand: News in Tweets", Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 42–51, Seattle, WA, USA, November 2009.
5. B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, M. Demirbas, "Short Text Classification in Twitter to Improve Information Filtering", Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10), pp. 841–842. New York, NY, USA, 2010.
6. F.P. Tellez, D. Pinto, J. Cardiff, P. Rosso, "On the Difficulty of Clustering Company Tweets", Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents (SMUC '10), pp. 95–102. New York, NY, USA, 2010.
7. T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, M. Dredze, "Annotating Named Entities in Twitter Data with Crowdsourcing", Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10), 2010.
8. M. Rowe, M. Stankovic, "Mapping Tweets to Conference Talks: A Goldmine for Semantics", Proceedings of the 3rd International Workshop on Social Data on the Web (SDOW 2010) at the 9th International Semantic Web Conference (ISWC 2010), 2010.