Twitter Sentiment Analysis of T20 World Cup 2016

Neha Janu
Swami Keshvanand Institute of Technlogy

Prashant Kathuria
Swami Keshvanand Institute of Technlogy

Rajat Gupta
Swami Keshvanand Institute of Technlogy

Tushar Mittal and Subroto Biswas
Swami Keshvanand Institute of Technlogy

***Abstract*—Conducting analytics over data generated by Social Web portals such as Twitter is challenging, due to the volume, variety and velocity of the data. Commonly, adhoc pipelines are used that solve a particular use case. In this paper, we generalize across a range of typical Twitter-data use cases and determine a set of common characteristics. Based on this investigation, we present our Twitter Analytical Platform (TAP), a generic platform for conducting analytical tasks with Twitter data. The platform provides a domain-specific Twitter Analysis Language (TAL) as the interface to its functionality stack. TAL includes a set of analysis tools ranging from data collection and semantic enrichment, to machine learning. With these tools, it becomes possible to create and customize analytical workflows in TAL and build applications that make use of the analytics results. We showcase the applicability of our platform by building Twindera search engine for Twitter streams.**

Keywords*—*Tweet Mining,Sentiment Analysis,Naive Bayes, Natural Language Processing,Event Analysis.

I. Introduction

Twitter, a popular microblogging service, has received much attention recently. It is an online social network used by millions of people around the world to stay connected to their friends, family members and co-workers through their computers and mobile phones. Twitter asks one question, What are you doing? Answers must be fewer than 140 characters. A status update message, called a tweet, is often used as a message to friends and colleagues. A user can follow other users; and her followers can read her tweets. A user who is being followed by another user need not nec- essarily have to reciprocate by following them back, which renders the links of the network as directed. After its launch on July 2006, Twitter users have increased rapidly. They are currently estimated as 44.5 million worldwide 1 . Monthly growth of users has been 1382 % year-on-year, which makes Twitter one of the fastest-growing sites in the world 2 .Some studies have investigated Twitter: Java et al. an- alyzed Twitter as early as 2007. They described the social network of Twitter users and investigated the motivation of Twitter users . B. Huberman et al. analyzed more than 300 thousand users. They discovered that the relation between friends (defined as a person to whom a user has directed posts using an @ symbol) is the key to under- standing interaction in Twitter. Recently, boyd et al. investigated retweet activity, which is the Twitter-equivalent of e-mail forwarding, where users post messages originally posted by others. Twitter is categorized as a micro-blogging service. Mi- croblogging is a form of blogging that allows users to send brief text updates or micromedia such as photographs or au- dio clips. Microblogging services other than Twitter include Tumblr, Plurk, Emote.in, Squeelr, Jaiku, identi.ca, and so on 3 . They have their own characteristics. Some examples are the following: Squeelr adds geolocation and pictures to microblogging, and Plurk has a timeline view integrating video and picture sharing. Although our study is applicable to other microblogging services, in this study, we specifically examine Twitter because of its popularity and data volume. An important common characteristic among microblog- ging services is its real-time nature. Although blog users typically update their blogs once every several days, Twit- ter users write tweets several times in a single day. Users can know how other users are doing and often what they are thinking about now, users repeatedly return to the site and check to see what other people are doing. The large num- ber of updates results in numerous reports related to events. They include social events such as parties, baseball games, and presidential campaigns,Big Sports Events etc.

A. Sentiment Analysis

This technology has many interesting consequences not just for businesses but to countries as a whole .For businesses, online opinion has turned into a kind of virtual currency that can make or break a product in the marketplace. On the other hand the aggregate of emotions can very effectively measure the temperature of a country in response to real world events like a mass increase in the level of worriedness around major weather phenomena a mass increase in the level of distress and sadness after terror attacks the same with political changes and so on. As sentiment analysis tools begin to take shape, they not only help businesses improve their bottom lines, but also eventually transform the experience of searching for information online. The simplest algorithms work by scanning keywords to categorize a statement as positive or negative, based on a simple binary analysis love evaluates to good, while hate to bad). But that approach fails to capture the subtleties that bring human language to life: irony, sarcasm, slang and other idiomatic expressions. Reliable sentiment analysis requires parsing many linguistic shades of gray. For example, several adjectives often signal a high degree of subjectivity, while noun- and verb-heavy statements tend toward a more neutral point of view. As sentiment analysis algorithms become more sophisticated, they begin to yield more accurate results that may eventually point the way to more sophisticated filtering mechanisms. It gives its users the ability to research any topic on blogs, social media sites, and in traditional news media reports. The rise of blogs and social networks has also affected the bull market (the stock exchanges) by reviews, ratings, recommendations and other forms of online expression. For computer scientists, this fast-growing mountain of data is opening a tantalizing window onto the collective consciousness of Internet users.

B. Tweet Analysis step

Figure 1 shows the steps performed to find the polarity of a tweet. This model is used for all the live tweets which are generated during 2016 T20 World cup. We have used NLP and Naive bayes classifier in this project and merged both of their result to find polarity of a tweet.

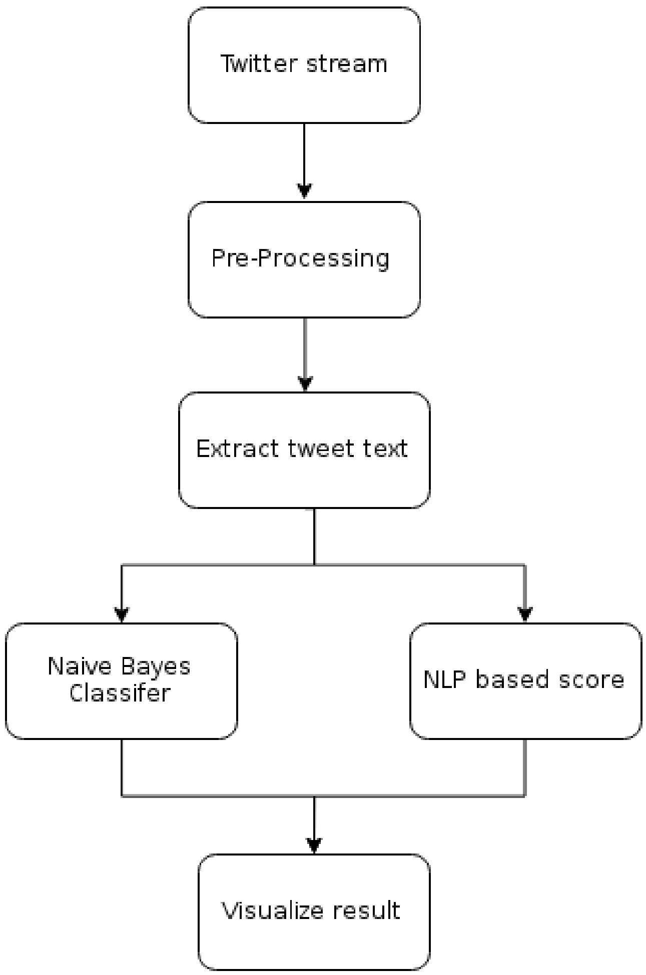

Figure 1:Flow Chart which depicts all the steps

C. Data Pre-Processing

As a first step towards finding a tweets sentiment and in order to obtain accurate sentiment classification, we need to filter some noise and meaningless symbols from the original text of tweets that do not contribute to a tweets sentiment. This was done by splitting up the text using spaces and constructing a bag of words, which is called tokenization. Each word can be used as a feature to train the classifier, but we needed to keep some prominent ones and remove some useless and meaningless words or symbols. Tweets may also include symbols; for example the word following the @ symbol is a username and is used to mark topics or keywords in a tweet. All usernames and URLs were converted to generic tags (e.g. all @usernames tagged as username), and some of mentions can be used to improve the performance of the sentiment classifier

We also extracted data like country, time , whether it is reTweeted or not and

D. Naive Bayes Classifier

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

$$\mathrm{p}(\mathrm{C}_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}$$

In plain English, using Bayesian probability terminology, the above equation can be written as

$$\text{posterior} = \text{prior} \times \text{likelihood} \frac{}{\text{evidence}}$$

E. NLP approach

The Natural Language Processing approach to sentiment analysis uses features provided by a NLP tool such as word sense disabmiguation, part-of-speach tagging(POS Tagger),Object Verb Subject Segmentation,Syntactic Parser etc. Using tweets dynamically collected from Twitter servers, sentiment orientation of other part of speech (noun and verb) is evolutionally estimated by the process sketched above, but also utilizing subject-predicate relationship. This way, the semantic orientation of words (Noun and Verb) is assessed by considering the semantic orientation of adjectives parts of speech which co-occurs as predicate with an already detected semantic orientation. It then extract each sentence from every document and parse it in the feature extractor which classifies it into the required sentiment category.

F. Team Support

In this paper we have focused on evaluating the support for each team playing the tournament using the following algorithm

```
TeamSupport(team1Set,team2Set,sentiment)
 For(each tweet)
  if(count(team1Set) > count(team2Set))
   If(sentiment=positive)
    Supports Team1
   else
    Supports Team2
  else
   If(sentiment = negative)
    Supports Team2
   else
    Supports Team1
```

Positive Tweets

- Amidst #AsiaCupT20Final hype, Faf du Plesis has played a sensational innings!!! #SAvsAus
- #AsiaCupT20Final The drainage in Sher-e-bangla is the best in the world The ground had puddles of water, not thr long back. Its all gone now
- Pitches accross d globe seem like interchanged.. That's d impression seeing T20s in SA n Aus and in Ban.. #SAvAUS #BANvIND #AsiaCupT20Final
- Good news from #Mirpur !!! Rain stopped and Groundsmen working in full strength, removing covers. #AsiaCupT20Final
- Mahmadulla is outstanding. It's a Treat to watch. :-) #AsiaCupT20Final

Negative Tweets

- This is boring hit some shots so India can chase in interesting way #IndvsSL
- No ball from Ashwin! :/ #INDvsSL
- What a boring match!!!! #INDvsSL
- With that run out of Shanaka, India has forced a run out in all of its matches in the Asia Cup this year so far. #INDvsSL #SLvsIND
- Bad luck or over confident umpire. #INDvsSL

Neutral Tweets

- Everyone start bitting ur nails this one gonna be edge #INDvsSL
- A fine inning comes to an end. Yuvraj 35 runs off 18 balls (3×6 3×4) #INDvsSL
- A solid knock with combination of aggression, maturity and stability @imVkohli #WellPlayedIndia #INDvsSL
- So till when will India keep persisting with Dhawan ? #INDvsAUS
- So Dhoni is not trying new players as they are not finished products.All The hard work done by top order will go down drain again #IndvsAus

Figure 2:Sample Training Data

II. Conclusion

As described in this paper, we analysed the sentiments of people on the basis of their tweets. Then using that data we visualized the polarity of tweets , Support of teams and which is more , analysed which platform is used more to tweet and location . We observed that the most no. of tweets were collected during the second semi final between India and West Indies. Twitter is known to be a negative platform and for having harsh opinions on every topic which was also observed in our project as most of the tweets collected were of negative sentiment. Microblogging platform such as twitter has real-time characteristics that distinguish it from other blogs and collaborative bookmarks.In this paper we presented an example using the real-time nature of twitter and the emotions of the general public during an international event.

References

[1] GOAALLL!: Using Sentiment in the World Cup to Explore Theories of Emotion Jonathan Gratch, Gale Lucas, Nikolaos Malandrakis, Evan Szablowski, Eli Fessler and Jeffrey Nichols

[2] Text Analysis and Sentiment Polarity on FIFA World Cup 2014 Tweets Peiman Barnaghi ,Parsa Ghaffari ,John G. Breslin

[3] A. Bifet and E. Frank, "Sentiment knowledge discovery in twitter streaming data," in Discovery Science, 2010, pp. 1-15.

[4] D. Terrana, A. Augello, and G. Pilato, "Automatic Unsupervised Polarity Detection on a Twitter Data Stream," in Semantic Computing (ICSC), 2014 IEEE International Conference on, 2014, pp. 128-134.

[5] Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.