Text and Sentiment Analysis of IPL (Indian Premier League) Tweets

Anirudh
Indiana University
Bloomington
MS in Data Science
aanirudh@iu.edu

Gautham Sriman Narayan
Indiana University
Bloomington
MS in Computer Science
gsrimann@indiana.edu

Vishwas Vijaya Kumar
Indiana University
Bloomington
MS in Computer Science
visvijay@indiana.edu

Abstract

As always, the world united this year in moments of triumph, activism, support, and fascination — and Twitter is where we gathered for all of it. Whether people were making a hashtag into a global movement or expressing wonder over a photo of a dress, we all used Twitter this year in awe-inspiring ways.

As TV watchers tweet about how they feel and what they see, they produce valuable information not only about the TV program but also how engaged they are to the program. This paper seeks to apply and extend the current work in the field of natural language processing and sentiment analysis to data retrieved from Twitter and apply sentiment analysis to determine public views about an event. We have classified data based on Paul Ekman's 6 emotions. Our dataset contains 60,000 tweets related to IPL.

The primary focus of this experiment is to check what sentiments these tweets reflect towards sports. A large collection of such sentiments could be leveraged to provide useful reflection of public sentiment towards sports event. Naive Bayes and support classifiers were used to tag given tweet as positive, negative or neutral. Bag of words, classifiers trained on Ekman's 6 emotions and Janyce Wiebe's subjectivity lexicon was used for polarity classifier.

1. Introduction

Live broadcast TV programs such as sports games, reality shows, and concerts often attract a large number of loyal watchers.

India is one of the fastest-growing markets for the company today because of the media partnership team's efforts to make Twitter more useful and relevant to all aspects of Indian society—from news to politics to sports to entertainment to TV. Stanton believes Twitter is the "social soundtrack to life.

One of the most popular topics among the Indians is IPL. IPL is full of excitement for everyone. IPL is not only popular in the boys like other games but girls are also interested in this tournament. IPL is a short form of cricket but it attracts people like a magnet.

The biggest T20 cricket league in the world saw cricket fans enthusiastically engaging in public conversation about the league with over 9 million #IPL Tweets last season. According to the Forbes @TwitterIndia just released the results of a self-commissioned survey which found that 89 percent of Twitter users in India are cricket fans. With extra support from networks like Star Sports, Set Max Twitter has become the go-to, second-screen app for fans in India. Campaigns and advertisements like #CrashThePepsiIPL make it easier for an ever-expanding social network, like Twitter, to uncover new markets and capitalize on those they already have.

The intent of this paper is to study behavior or attitude of users towards IPL

(Indian Premier League). We chose twitter as medium for following reason:

(a) Tweets posted by TV watchers directly reflect what they feel and think about the game. Twitter analysis directly allows us to learn about the program audience. With over 300 million active users, Twitter ensures good coverage of popular TV programs.
(b) Public tweets are free and are easy to retrieve due to their brief, textual nature.
(c) The textual nature of tweets makes them amenable to lexicon-based analysis. As a result, end users can easily personalize Twitter-based event recognition and sentiment extraction by using the right keywords.
(d) Finally, Twitter allows tweets to be retrieved by tweeters.

To demonstrate the feasibility of using Twitter for Sentiment Analysis, we choose Cricket as sports and Indian Premier League or IPL as topic for our demonstration.

In this workshop paper, we present our ongoing work in extracting audience sentiments from Twitter analysis. Although we focus on IPL and Cricket in this work, most of the techniques can be readily applied to many other sports games that have a similarly sized fan population and have similar frequencies of major events, e.g., soccer, baseball, and basketball.

The rest of the paper is organized as follows. Section 2 talks about related work which motivated us and how it was relevant to our topic. Section 3 provides detailed explanation about Data. Section 4 provides detailed explanation about techniques and approach used. Section 5 demonstrates the effectiveness of our approach followed by Discussion & Conclusion and References.

2. Related Work

Several concurrent projects also study tweets about sports games, however they do not provide real-time event detection. Hannon et al [1] used post rate of tweets to produce video highlights of the World Cup offline. They did not recognize game events nor did they produce highlights in real-time. Chakrabarti and Punera [2] assumed that a game event is already recognized and focused on describing the event using Hidden Markov Models trained with tweets collected from events happened in the past. Therefore, our focus on real-time event recognition is complementary, and addresses a more difficult and fundamental problem.

Bollen et al. [3] modelled public mood and emotion according to people‟s Twitter posts. Pandey and Iyer, Barbosa and Feng [4] proposed machine learning approaches to classify sentiments on tweets. They focus on tweets with certain expressions over a long time, e.g. one year in [3].

Ekman's [4] point us to what we need to learn about emotions and how we can classify them into different categories. Mohammed and Turney [6] showed how emotion detection can be used as a tool for social and literary analysis, also showed that how combined strength and wisdom of the crowds can be used to generate a large term emotion association lexicon.

Mihalcea and Liu [7] have focused in their work on two particular emotions – happiness and sadness. They work on blog posts which are self-annotated by the blog writers with happy and sad mood labels. Saima and Stan [8] describe emotion annotation task of identifying emotion category, emotion intensity and word/phrases that indicate emotion in text which proved to be very useful. Ritter, A., Cherry, C., & Dolan, B [13] presented an approach that allows the unsupervised induction of dialogue structure from naturally-occurring open-topic conversational data which was very valuable for this experiment. Mageed and Diab [14] helped us understand subjectivity and sentiment annotation in natural language.

Existing works on sentiments measurement and opinion detection focus on product review and tweets moods modelling. Hu and Liu [9] mined and summarize the customer reviews of a product. Pang et al [10] and Zhuang [11] focus on sentiments classification in movie reviews. Jansen et al [12] investigated Twitter as a form of electronic word-of-mouth for sharing consumer opinions concerning brands. Extracting sentiments from sport games-related tweets are significantly different from product or movie reviews, because reviews have formal format and rich context information but tweets are colloquial without context. Furthermore, as people are emotional during sports games, their sentimental expressions are diverse and unexpected. Gratch ,Lucas ,Malandrakis, Szablowski, Fessler and Nichols [15] helped us how similar experiment was conducted in different sports. Sporting events can serve as natural laboratories for understanding how people emotionally respond to situations. Sport often involves high stakes and certainly evokes strong emotions in both participants and observers.

3. Data

Twitter: - The IPL data was gathered from twitter.com website. The first task is to build a data set. Twitter provides APIs that we can use to interact with their service. We used the 'Tweepy" library for this purpose.

Tweepy: - is a Python 2.6, 2.7, and 3.x library for accessing Twitter. It provides access to all twitter RESTful API methods, including reading and posting of tweets. Tweepy supports OAuth authentication, as BasicAuth is no longer supported by the Twitter API.

Both the Twitter search API and streaming API are available in tweepy**.** The Twitter streaming API was made use of to download the tweets in real time. In tweepy, an instance of 'tweepy.Stream' establishes a streaming session and routes all messages to the 'StreamListener' instance. The 'on_data' method of a stream listener receives all messages and calls functions according the message type. The Streaming APIs give developers low latency access to twitter's global stream of Tweet data. Connecting to the streaming API requires keeping a persistent HTTP connection open.

We ran the program for 3 days to gather tweets with hastags #IPL, #ipl2016. We chose English as our language of preference so that we get relevant tweets which can be used for sentiment analysis. We collected close to 60,000 tweets after running the program for 3 days which was close to 250 MB of Data. One of the reasons for collecting large number of tweets was that there are 8 teams involved in tournament and less number tweets might result in biased results in favor of particular team.

4. Methodology

In the implementation of our project, we embraced a 'semi-supervised' machine learning approach, as we had a larger proportion of unlabeled data compared to the labeled data. Broadly, we accomplished the following tasks:

1) Collect about 60,000 tweets relating to our event of interest, the "Indian Premier League (IPL)".
2) Inferred the trending topics related to the IPL.
3) Developed a classifier to perform 'Sentiment Analysis' of the twitter data based on Ekman's Six emotions.
4) Determined the 'Polarity' of the tweets, by classifying the same into 'positive', 'negative' and 'neutral' categories.
5) Visualized the results to make it better comprehensible.

4.1 Pre-Processing

When we collect the tweets, they may be ‘noisy’ in their original form. Many Twitter users use clumsy, ungrammatical sentence structures, and we would have to weed out the irrelevant parts of the tweet, like the URLs.

We did this ‘data cleanup’ of the twitter data by tokenizing the tweets and making use of a list of ‘stopwords’ (and removing the same). We also removed the punctuations that had little or no semantic meaning towards a tweet.

4.2 Classification

For the classification of our twitter data, we used the Naïve Bayes Classifier.

Naïve Bayes: Naïve Bayes is a simple yet effective probabilistic technique to construct a classifier that is based on the Bayes’ theorem. The Naïve Bayes classifier makes an assumption that the presence of a specific feature of a class is in no way related to the presence of any other feature.

Naïve Bayes models use parameter estimation pertaining to the method of ‘maximum likelihood’. These classifiers typically outperform the other classifiers when the sample sizes are small. [22].

The Bayesian probability can be represented as [23]:

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

This can be re-written as follows [23]:

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \; p(C_k) \prod_{i=1}^{n} p(x_i|C_k).$$

Where, the above equation is a Bayes classifier which is basically a function that performs the task of assigning ‘y’=C (a class label) for some ‘k’.

We used 3 lexicons for our classifiers, namely ‘NRC Emotion lexicon’, ‘Janyce Wiebe’s subjectivity lexicon’ and ‘Ekman’s word-emotion lexicon’.

4.3 Features

We used a “Bag of Words” model for this purpose.

Bag of Words: In a Bag of Words model, the representation of text is in the form of a word-multiset, where we don’t pay any heed to the order or the grammatical structure, but we take the multiplicity into consideration. This model is widely used in text/document classification methods, where the word frequency is typically utilized as a ‘feature’ for classifier training. Hence, the bag of words model is an important form of information representation in areas like Information retrieval and Natural Language Processing (NLP) [23].

4.4 Visualization

For the visualization of our results, we used R. We generated word clouds to depict the ‘trending’ topics and the ‘emotion’ classification (based on Ekman’s 6 emotions). We also made statistical graphs for the ‘Sentiment Analysis’ in order to show the ‘Polarity’ of the emotions of the sports tweets we gathered, and classified the same into 3 categories- positive, negative and neutral.

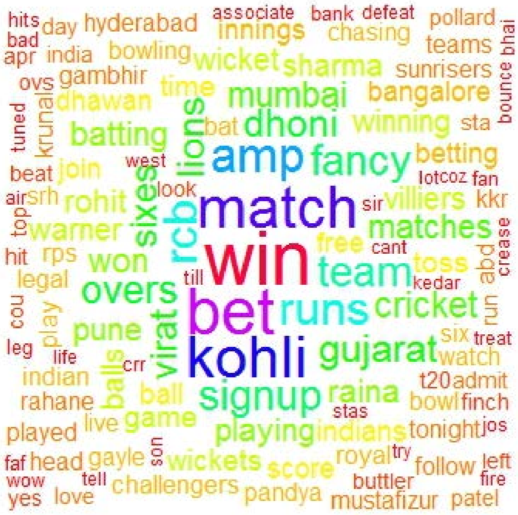


Figure 1: A ‘Trending’ word cloud

5. Evaluation

We cleaned 10,000 of the most recent IPL-related tweets for the purpose of evaluation. From this corpus of tweets, we were able to find out (by using Ekman's 6 emotions categorization) that there were ~1000 'joy' tweets, ~500 'anger' tweets, ~300 'surprise' tweets, ~200 'sadness' tweets, ~100 'fear' tweets, and about 8000 tweets that were classified as 'unknown'.

This shows that most of the sports-related tweets are factual and display comparatively less emotional words. This is quite typical of people who follow Sports, because they are usually more inclined towards scores, statistics and factual information, and pay less attention to the emotional parameters while expressing their point of view on social media and other platforms.

Below, we give you the graph of the results depicting the emotion categories:

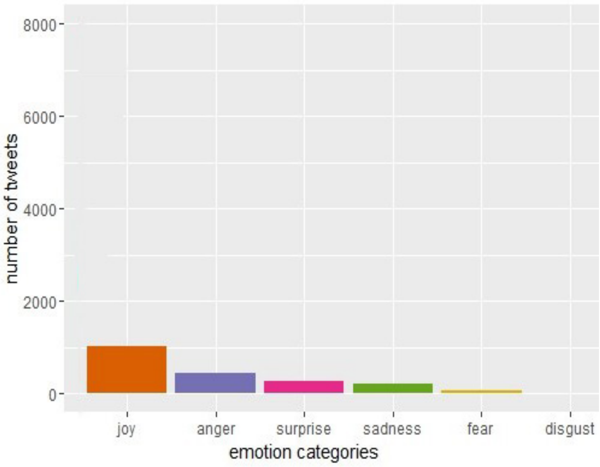


Figure 2: Graph showing the emotions towards IPL

We also determined the 'Polarity' of the most recent 10,000 IPL tweets, and found out that about 7,200 tweets were 'positive', 1,800 tweets were 'negative', and about 1,100 tweets were classified as 'neutral'.

Below, we have given examples of some of the tweets that were classified according to the 'Polarity':

Positive
• Memorable night for Knight Riders: Gambhir and Uthappa come good yet again; Russell chips in with four wickets • Cutest Moment You'll Ever See In An #IPL Match. SRK With AbRam.
Negative
• #Kxip 11 has 6 out of form players they need to sort out things miller saha etc #KKRvKXIP #KXIPvKKR #ipl \n#IPL2016" • IPL 2016, DD vs RPS: Pune's IPL in danger of early finish as MS Dhoni struggles to get a grip - The Indian Expr...
Neutral
• Delhi pitch and association keep changing: Bhatia • Captain of @RPSupergiants @BCCI and former @ChennaiIPL is ready for tomorrow #RPSVDD #IPL2016 #IPL

The results of the 'Polarity' classification are depicted in the following graph:

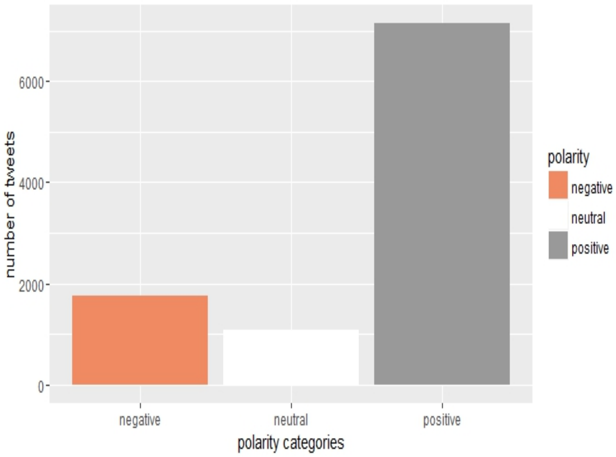


Figure 3: Graph showing the polarity of people towards IPL

6. Conclusion

The most popular communication tool to share everyday opinions and life events is social media. Twitter is one such major online social networking service. We can analyze the sports-related tweets in real time, and infer meaningful relationships. When aggregated, these tweets reflect public sentiment towards sports. We can extract the sentiment from them and look at the general correlation between these sentiments and a sports event. A large collection of such tweets could be leveraged to provide a useful reflection of public sentiment towards sports.

As we have shown, we can perform a very accurate Sentiment Analysis of Sports events. In our case, we had chosen India’s largest sports league, the “Indian Premier League (IPL)”. This same kind of analysis can be applied in a broader way in other Sports as well.

We could have gathered data from other Social Media platforms as well. apart from Twitter, and, this is what we intend to do in the future to expand our work.

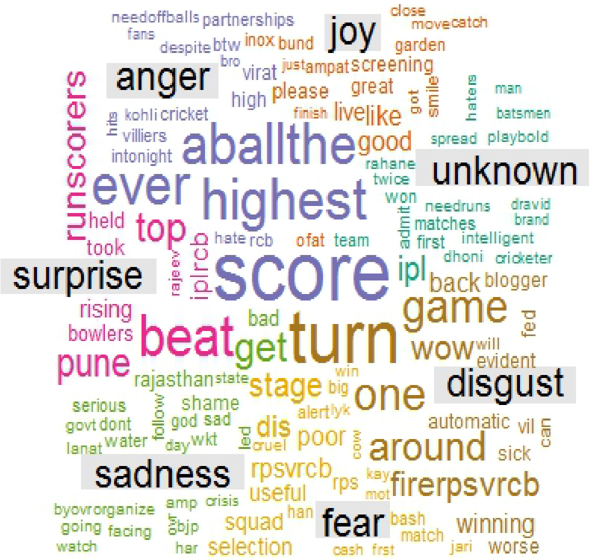


Figure 4: An ‘Emotion’ word cloud

References

1. J. Hannon, K. McCarthy, J. Lynch, and B. Smyth, "Personalized and automatic social summarization of events in video," in Proc. ACM IUI, 2011.

2. D. Chakrabarti and K. Punera, "Event Summarization using Tweets," in Proc. AAAI ICWSM, 2011.

3. J. Bollen, A. Pepe, and H. Mao, "Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena," in Proc. AAAI ICWSM, 2011.

4. L. Barbosa and J. Feng, "Robust Sentiment Detection on Twitter from Biased and Noisy Data," in Proc.ACM COLING, 2010

5. Paul Ekman's An Argument for Emotion, University of California ,San Francisco, U.S.A

6. Saif M. Mohammad and Peter D. Turney Crowdsourcing a Word–Emotion Association Lexicon Saif M. Mohammad and Peter D. Turney Institute for Information Technology, National Research Council Canada.Ottawa, Ontario, Canada, K1A 0R6

7. Mishne, G., Glance, N.: Predicting Movie Sales from Blogger Sentiment. In: AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (2006)

8. Saima Aman and Stan Szpakowicz, Identifying Expressions of Emotion in Text ,1 School of Information Technology and Engineering, University of Ottawa, Ottawa, Canada 2 Institute of Computer Science, Polish Academy of Sciences, Warszawa,
Poland

9. M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proc. ACM SIGKDD, 2004.

10. B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in Proc. ACM EMNLP ACL-02 Volume 10, 2002.

11. L. Zhuang, F. Jing, and X.-Y. Zhu, "Movie review mining and summarization," in Proc. ACM CIKM, 2006

12. B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as

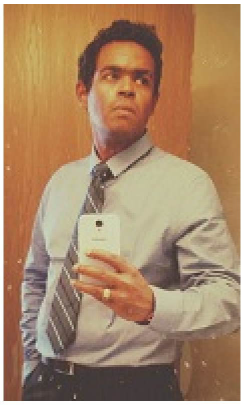
electronic word of mouth," Journal of the American Society for Information
Science and Technology, vol. 60, pp. 2169-2188, 2009.
13. Unsupervised Modeling of Twitter Conversations Ritter, Alan; Cherry, Colin; Dolan, Bill
14. Muhammad Abdul-Mageed and Mona T. Diab, Subjectivity and Sentiment Annotation of Modern Standard Arabic Newswire
15. GOAALLL!: Using Sentiment in the World Cup to Explore Theories of Emotion by Jonathan Gratch, 1 Gale Lucas, 1 Nikolaos Malandrakis, 2 Evan Szablowski, 3 Eli Fessler4 and Jeffrey Nichols
16. Analyzing Twitter for Social TV: Sentiment Extraction for Sports by Siqi Zhao and Lin Zhong and Jehan Wickramasuriya and Venu Vasudevan.
17. https://blog.twitter.com/2015/the-2015-yearontwitter-in-india-in
18.http://www.sporttechie.com/2015/02/19/twitters-strategy-to-capitalize-on-the-cricket-world-cup/
19.http://www.livemint.com/Consumer/530pe30yF3C56xPJEIEUMM/Cricket-on-Twitter-is-being-driven-globally-from-India-Kati.html
20. http://www.oceanofweb.com/cricket/ipl-indian-premier-league.html
21.http://www.sporttechie.com/2015/02/19/twitters-strategy-to-capitalize-on-the-cricket-world-cup/
22. www.ic.unicamp.br- "Naïve Bayes Classifier"
23. www.wikipedia.com

About the authors

1) Anirudh

Current Master's student in Data Science at the School of Informatics and Computing. I have more than 5 years of work experience in QA and testing and have worked for companies like EA Games, New York Times. My main areas of interest are Statistics, Data analysis, Data mining and Machine learning. Even though I don't like social media so much but you can find me at
Facebook:
https://www.facebook.com/pillaianirudh
LinkedIn:
https://www.linkedin.com/in/pillaianirudh

I missed one class in March because I wasn't feeling well.



2) Gautham Sriman Narayan

Current Master's student in Computer Science at the School of Informatics and Computing. My main areas of interest are Big Data Analytics and Security.

You can find me at:
www.facebook.com/gautham.sriman

I missed one class in April due to illness.

3) Vishwas Vijaya Kumar

Current Master's student in Computer Science at the School of Informatics and Computing.

My main areas of interest are Cloud Computing and Data Mining.

You can find me at:
Facebook:
www.facebook.com/kumarvis.58?fref=ts

I missed two classes, the first one due to an interview and the second due to a conference