International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Real Time Tweet Summarization and Sentiment Analysis of Game Tournament

Vikrant Hole[1], Mukta Takalikar[2]

[1]Pune Institute Of Computer Technology, Department Of Computer Engineering, Pune, India

Abstract: ***Streaming data analysis in real time is becoming the fastest and most efficient way to obtain useful knowledge from what is happening now, allowing organization to react quickly when problem appear or to detect new trends helping to improve their performance. There is substantial volume of tweet about each event, and there is press coverage of each event to serve as a gold standard. The problem we try to solve is that in absence of being live in front of television set, build data processing system that create informative data regarding tournament matches and correlate sentiment of fans to match play. Using twitter data, we find sub-events in game and sentiment of fans posting tweet related to game. We propose a system for real time summarization of scheduled sub-events for games using twitter data. We also propose an approach that analyzes feelings of people posting tweet. We focused on summarizing sporting events, specifically FIFA World Cup 2015 and IPL 2015. For a system using social media like twitter to keep track of things happening around, we look for following traits: (I) Detection of bursty topic as soon as it emerges; (II) Summarization of related bursty topic; (III) Analyzing feelings of fans and correlating sentiment to game.***

Keywords: summaization, sentiment analysis,big data.

1. Introduction

Microblogging today has become a very popular communication tool among Internet users. Millions of messages are found on social networking sites like twitter1, facebook2, tumblr3 which could be used marketing and social studies. Study of the user data of social networks is one of the current trends of the times. Big data can be used for collection of large and unstructured data which is difficult while using traditional database. Handling the complexity of big data it is possible to use it for understanding data pattern and use it for learning to predict.

The tweets that are done may be related to different topics and the business man want to know more about it. But the vastness of data over twitter restricts them from undergoing overview. So the need of summarization is there which will provide a better solution.

We use data collected from twitter which is in form of messages. The content of messages varies from personal to social views. Natural Language opinions are expressed in restrained and multifarious ways, which are difficult to solve by basic text processing methodologies. Recognizing the sentiment and sub-events correctly is more tedious due its unrestricted message format.

Sentiment analysis is detection of attitude enduring, affectivity colored beliefs, disposition towards person or object. Sentiment consist of source i.e the holder of sentiment, target to whom it is going to affect and type of attitude i.e loving, hating, value, desire and etc. typically the typology of affective states are emotion, mood, interpersonal stances, attitude and personality traits according to Scherer.

In the proposed system, we perform Real-Time summarization of scheduled sub-events for game tournaments like soccer and cricket from twitter stream. We propose an approach that analyses feeling of soccer and cricket fans and correlate sentiment to match play. We focus on summarizing sporting events, specifically World Cup soccer matches, because each event takes place over a short defined period of time, there is a substantial volume of tweets about each event, and there is press coverage of each event to serve as a gold standard. For a system using social media like Twitter to keep track of things happening around, one would be looking for the following traits: (I) Detection of a busty topic as soon as it emerges; (II) Summarization of related busty topics; (III) Analyzing feeling of soccer fans and correlate sentiment to match play.

2. Literature Survey

In our proposed system, we perform real-time summarization of scheduled sub-events for game tournaments like soccer and cricket from twitter stream. We also propose an approach that analyses feeling of soccer fans and correlate their sentiment to match play. This is the first work according to the survey done by us and so we need to do literature survey into two section one consisting of Real-Time summarization of scheduled sub-events for game tournaments and second sentiment analysis.

A. Literature Survey of Real-time Summarization

In Real-Time summarization of tweets, we use soccer dataset which contain tweets posted by fans reporting sub-events in match. There are two operations to be performed in Real-Time summarization. First, finding important moments during match like sub-events expressing goal, passes, shots etc. during match. Second, finding few important tweets which best express the identified sub-event cluster. So redundancy in dataset is reduced. Chakrabarti et al. [1] used Modified Hidden Markov Chain to learn structure of events and used football match tweets held at America. But it did not identify all the sub-events. As it used unsupervised approach, it was not possible to identify participants and sub-events unless match was played multiple times between same two teams. Chen lin et al. [2] used Dynamic pseudo relevance feedback (DPRF) language model for relevant tweet in event and Graph based optimization for temporal

continuity and content coherence in storylines. It detected Query specific event.

Nichols et al.[3] summarized sporting event using an unsupervised algorithm for generating a textual summary of events from status updates in Twitter. Sub-event detection (Moments) was detected by assumption that rapid increase in status update indicates important sub-events. Selecting Moment (Summary) TF-IDF and Sentence Scoring Approach were used. Shen et al.[4] detects the important sub-events related to each participant and finally summarize events. It used Gaussian Distribution for time aspect of topic and multinomial distribution for content aspect of topic. It used TF-IDF for summarization

Arkaitz et al. [5] proposed, two-step process for the realtime summarization of events sub-event detection and tweet selection, and analyze and evaluate different approaches for each of these two steps. Sub-Event Detection was done evaluating increase in spike and outlier in spike (whether the tweeting rate for a given time frame stands out from the regular tweeting rate) Tweet Selection (Summary) was done on the basis of score each tweet which is the sum of the values of the terms that it contains.

Mitsumasa et al. [6]proposed method that generates sport update finding good reporter. Sub-Event Detection was done by detecting Burst in tweet and Tweet selection was done using tweet score and Reporter score combination. Corney et al. [7] proposed an approach for Fan finding and sub event detection by identifying word or phrase that show sudden increase in frequency. Finding co-occurring words across message to find event in burst and Fan identification was done using for each user, we count the total number of times they mention each team across all their tweets.

B. Literature Survey of Sentiment Analysis

Several research works have been carried out for social network analysis and sentiment analysis for deriving mood of people with respect to some product. Sentiment analysis study has been carried over decades. Is review positive or negative, these classification is done using only words(tokenization), i.e. cool is positive and disappointing is negative using polarity to create wordnet of positive and negative sentiment[10]. Sentiment tokenization issues like handling HTML and XML markup, Twitter markup(names, Hashtags), Capitalization, Numbers are difficult to handle, however approaches like utilizing n-grams, Part-of-speech tagging have been employed effectively in [8] for finding the twitter sentiment using the machine learning techniques and other methodologies.

Different sentiment lexicon providing different classes of positive, negative, strong, pronoun, quantifier and many more have been used to create wordnet of positive and negative sentiment lexicon[13]. SENTIWORDNET is the result of the automatic annotation of all the synsets of WORDNET according to the notions of positivity, negativity, and neutrality. Each synset s is associated to three numerical scores Pos(s), Neg(s), and Obj(s) which indicate how positive, negative, and objective (i.e., neutral) the terms contained in the synset are[12]. Earlier only words(tokens) were used for sentiment analysis and using it was not possible to identify towards whom the sentiment is intended. Inorder to find correct sentiment of sentence it is important for finding aspect or attribute target of sentiment. For example, Food was great but Service was bad, here there is one sentiment for one attribute while other for other one called micro sentiment. There are two approaches for it. First consist of phrases and rules where we find frequent frequency phrases i.e. fish, paneer etc. then filter these by rules like occur right after sentiment word i.e. great fish. Second find aspect in advance and find dataset related to it have been proposed by Jiang et al.[17]

Johan Bollen et al.[14] used POMS score to establish the sentiment values classified data as positive or negative in moods category using a syntactic, term-based approach, in order to detect sentiment. This method does not use any Machine learning algorithm for training but uses term based approach to classify sentiment.

Yong-soo et al.[15] proposed the novel approach to utilize situational information and personality of emotional subject. To extract and utilize situational information, it propose to use situation model using lexical and syntactic information. In addition, To reflect personality of emotional subject, it propose personalized emotion model using KBANN (Knowledge based Artificial Neural Network). Anjaria et al.[9] proposed hybrid approach of extracting opinion using direct and indirect features of Twitter data based on Support Vector Machines (SVM), Naive Bayes, Maximum Entropy and Artificial Neural Networks. He also used PCA with SVM in an attempt to perform dimensionality reduction.

Support Vector Machine (SVM), NaIve Bayes (NB) and Maximum Entropy (MaxEnt) classifiers are well discussed in many literatures such as Pang and Lee [11][16][17], whereas Artificial Neural Networks (ANN) have been discussed very limited number of times. Rodrigo Moraes et al. [16] discussed the comparative features of ANN and SVM in detail for the document level sentiment classification.

3. Methodology

In the proposed system, we fetch the tweets using Twitter API v 1.1. In order to remove stop words and extract features, we perform data cleaning and normalization. We search using twitter search API v 1.1 to collect data with various hash-tags like #BRAZIL, #USA, #FIFAWORLD CUP for collecting tweet related to match. The collected tweets have to undergo various stages to identify sub-events and sentiment of fans during match play as shown in Figure 3.1. In following sub-section we will describe each stage broadly.

This work presents a NEW EVENT DETECTION (NED) system that operates on-line with social streams. The objective was to design a system capable of detecting all events, and not just those which caused a significant spike in document volume. NED was designed to operate with high volume social streams. These streams necessitate the use of various heuristics to ensure computation is feasible. Therefore, the NED architecture was modulated into three

parallel processes: Stream Processor, Cluster Manager and Event Recognizer.

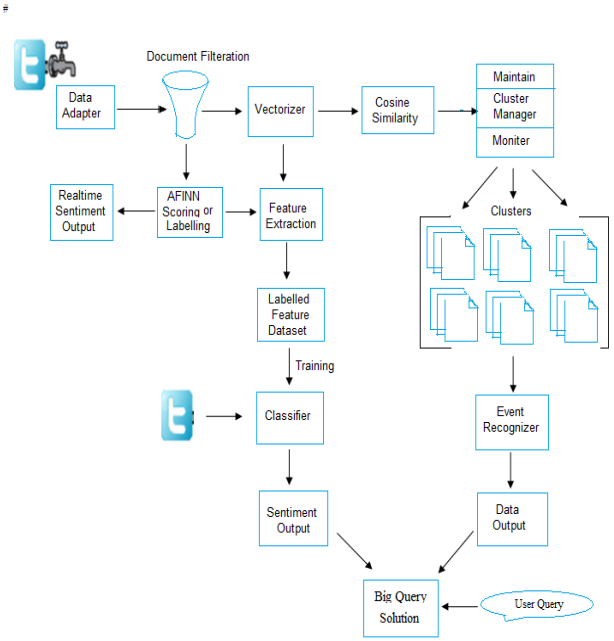

Figure 3.1: System Architecture

This work also presents identifying sentiment of fans during match with respect to team. We use a hybrid approach of extracting opinion using direct and indirect features of Twitter data based on Support Vector Machines (SVM), Naive Bayes, Logistic Regression and AFINN scoring. Figure 3.1 shows complete architecture of system.

A. Preprocessing

1. Data Adapter

The Data Adapter component is the link between the social stream and NED. This component was designed to retrieve one document at a time from the stream in chronological order. This document is then passed on to the Document Filtration component.

2. Document Filtration

The noisy nature of social streams necessitates a filtration layer. The Document Filtration component was designed to provide such a layer. Documents which aren't event-related are filtered out before providing it for sub-event identification. This design eliminates further computation on irrelevant documents. The Document Filtration component receives the documents from the Data Adapter and passes them on to the Vectorizer if they pass the filter test. This design also removes hyperlinks, username etc. before providing it to vectorizer. The implementation of the filter test will vary depending on the social stream.

3. Document Similarity and Clustering

Clustering algorithms are more efficient approaches to the NED task. These algorithms create clusters of documents with the goal that the documents in a cluster all report on the same event. When classifying a new document, these algorithms compare the centroids of the clusters to the new document. If the distance is within a certain threshold then that document is added to that cluster. Otherwise it is classified as NEW and a new cluster is created. Algorithm 1 shows a basic clustering approach.

The majority of NED applications under TDT use clustering algorithms. These algorithms use the tf.idf term weighting formula with the cosine similarity metric to determine cluster membership. A training phase is required to determine the threshold t and the initial idf values [6].The state of the art in cluster algorithms differ in their use of the tf .idf weighting.

For example, Allan et al. used a modified version of tf.idf weighting that factored in document age [7]. Brants et al. used an incremental tf.idf model where they updated the idf values periodically [10].

Algorithm 1 Clustering algorithm for the NED task

Require: t ⟵ input threshold
1: for all d in documents do
2: dis_{min},c ⟵ min_c { distance(d, clusters) }
3: if dis_{min} > t then
4: d ⟵ NEW
5: create new cluster(d)
6: else
7: d ⟵ OLD
8: c ⟵ d

This approach allows new terms to be added to the vocabulary as they appear in new documents. Various other researchers have made slight modifications to the tf.idf model. Schinas et al. boosted the weights of terms they deemed relevant to the events they were detecting [11]. They used a key-word lexicon based approach to detect these words. Makkonen et al. and Yang et al. proposed new vector models where multiple vectors were constructed per document [12, 13]. These vectors represented separate semantic classes, and they computed similarity between documents by comparisons in their vector sets.

B. Event Detection

Stream Processor consumes one document1 at a time from the social stream. This process was designed to represent each document in vector space and compute its nearest neighbor. The tuple of document and nearest neighbor is then sent to the Cluster Manager process.

1. Cluster Manager

Cluster Manger maintains and monitors clusters of similar documents. This process receives each new document and its nearest neighbor from the Stream Processor process. The decision is then made whether to assign the document to an existing cluster or create a new one. This decision is made based on the similarity between the document and its nearest neighbor. Events are detected by monitoring the growth rate of the clusters. Once an event has been detected, the corresponding cluster is sent to the Event Recognizer process. The Cluster Manager process is designed to perform two functions: to maintain clusters and to monitor clusters. The central component to this process is the Cluster.

2. Event Recognizer

The final process in the NED system is called Event Recognizer. This process receives clusters from the Cluster Manager process that are classified as new events. The role of this process is to recognize these events given knowledge of the structured events the system is trying to detect. This process was designed to include knowledge of the events possible during the structured events. Using this knowledge, the system attempts to categorize each cluster into a possible event. If successful, this event is output from the system. Otherwise, the cluster is rejected. This design helps alleviate some false positives from the previous processes. This process results are forwarded to event classifier.

3. Event Classifier

A keyword lexicon-based approach was taken. The design assumes a list of predefined events is available, and a list of keywords or phrases for each event. For example, a possible event may be `goal' and the list of keywords associated with that event could be [`score', `scores', `scored', `goal', `goals']. The algorithm for classifying an event is shown in Algorithm 2.

Algorithm 2 Event Classification

```
1: for all doc in cluster do
2:         for all event in predefined events do
3:                 lexicons  get_lexicons(event)
4:                 if any lexicon in doc then
5:                         votes[event] ++
6: event  ←  max(votes)
7: if event.count > (cluster.count=2) then
8:         output event
9: else
10:        output unclassified
```

The classification algorithm was designed as a voting system. Each document in the cluster increments the vote for the predefined events that it contains the keywords of. The predefined event that the cluster is classified as is that which has the most amounts of votes after this algorithm. A final check is applied to ensure that over 50% of the documents in the cluster voted for that event. The Data Output component was designed to output the events detected by NED.

C. Sentiment Analysis

Fans express their feeling and through sentiment analysis we detect the sentiment targeted towards any player or team with respect to any aspect like passing, goal, kick and etc. We use a hybrid approach of extracting opinion using direct and indirect features of Twitter data and classify sentiment using Support Vector Machines (SVM), Naive Bayes, Logistic Regression and sentiwordnet.

1. Sentiment Classification using Sentiwordnet

Opinion lexicons are resources that associate sentiment orientation and words. Sentiwordnet AFINN list is used. The term's glosses are then used to train a machine learning classifiers. Sentiwordnet is mainly used in real time sentiment classification as there is no training set. It is also used for labeling dataset that is to be used by machine learning classifier.

2. Sentiment Classification using Machine Learning Techniques

Sentiment Classification using Machine Learning Techniques require a model that can be used for classifying sentiment. Model is developed using training dataset. After developing model it is used to further classify the testing dataset. We have used NaIve Bayes, SVM, Logistic Regression classifiers with features extracted from Twitter data using feature extraction methods for sentiment analysis. We fetch the tweets using Twitter API v 1.1. In order to remove stop words and extract features, we perform data cleaning and normalization. We extract the target based extended features model [17] by modifying it and twitter user data from the normalized data. This feature vectors are used in part of chunks to train the classifier as a part of incremental training. After utilizing nearly half of the data we test it with half of the data.

4. Dataset and results

A. Dataset

The System is tested on tweets from five different matches. IPL2015 qualifier 1, eliminator, qualifier 2 and final match tweets used. The system was also tested on FIFA2015 final match. The tweets were fetched using #ipl and #UCLfinal hashtags.

B. Event Detection

Sporting events consist of a sequence of moments, each of which may contain actions by players, the referee, the fans, etc. At a high level, nearly every algorithm relies on spikes of the Twitter stream and we also use the same. Sudden increases, or "spikes," in the volume of tweets in the stream suggest that something important just happened because many people found the need to comment on it. Figure 4.1 shows spikes in tweets during game.

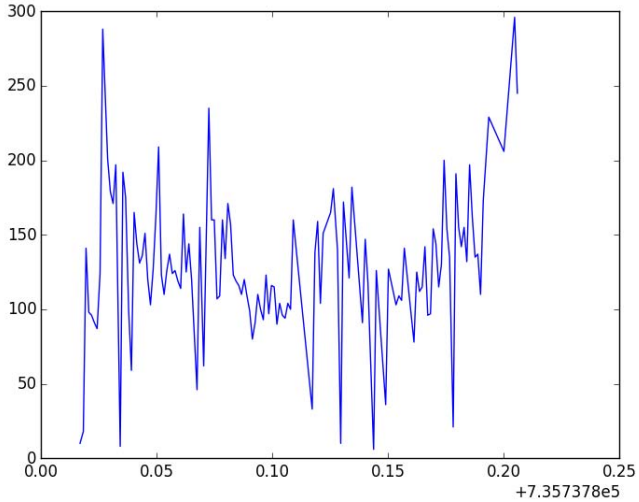

Figure 4.1: Number of tweet VS Time

The system is able to recognize sub-events in games with nearly 85%. In cricket it was able to identify sub-events like wicket, six, four, half-time and etc. In football, it identified sub-events like goal, penalty, half-time, red-card, yellow-card and corner. Table 4.1 show events and players rank according to tweets for qualifier 1 in IPL2015 match.

Table 4.1: Qualifier1 Rank

Rank	Token wise			
	Player		Event	
	Name	Tweet Count	Name	Tweet Count
1	Dhoni	937	Out	859
2	Bravo	874	Run	798
3	Nehra	574	Over	661
4	Parthiv	414	Wicket	534
5	Pollard	392	Ball	372

C. Sentiment Analysis

The system used SVM, NaïveBayes and Logistic Regression classifier for classifying sentiment of fans. After comparing the results, we conclude that SVM is able to classify tweets sentiment with higher precision but requires more time for training and testing. Figure 4.2 shows precision of different classifier for different matches. Table 4.2 shows the time required for training and testing different classifier in detail..

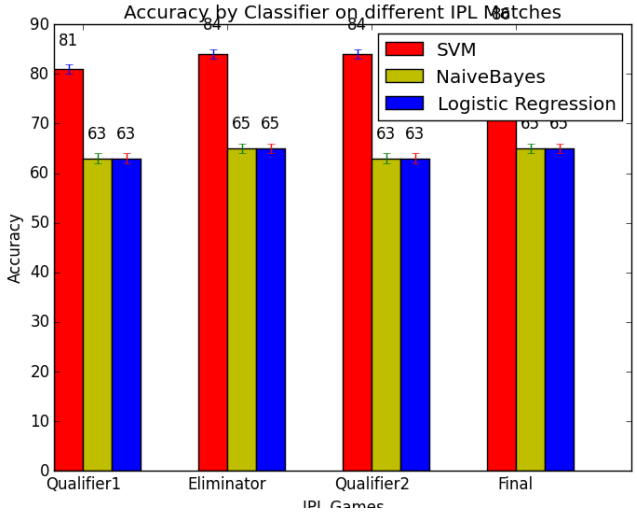

Figure 4.2: Sentiment Classifier Accuracy

Figure 4.3 shows pie chart to represent sentiment of Mumbai Indian and Chennai SuperKings fans for qualifier 1 in IPL2015. Figure 4.4 plots sentiment and sub-event in qualifier1. Green lines represent sentiment of Mumbai Indians while blue represent sentiment of Chennai Superkings. The red circles point wicket sub-event in game.

Table 4.2: Sentiment Classifier Time Analysis

IPL Game	*Dataset No. of tweet*	*Machine Learning Classifier(Time in seconds)*		
		SVM	*Naïve Bayes*	*Logistic Regression*
Qualifier 1	Train (8000)	6.193	0.31	0.171
	Test	4.228	0.99	0.102
Eliminator	Train (8000)	6.84	0.29	0.156
	Test	3.19	0.65	0.02
Qualifier2	Train (8000)	6.8	0.28	0.17
	Test	4.66	0.983	0.015
Final	Train (8000)	6.58	0.28	0.17
	Test	5.85	1.29	0.015

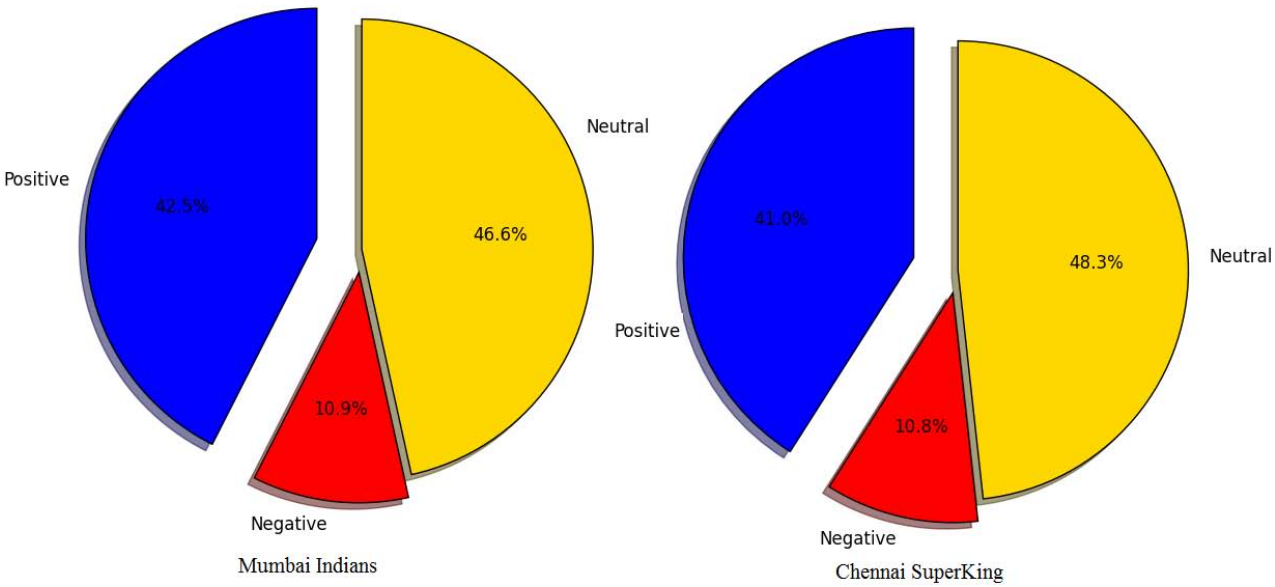

Figure 4.3: Team wise Sentiment Analysis for Qualifier1

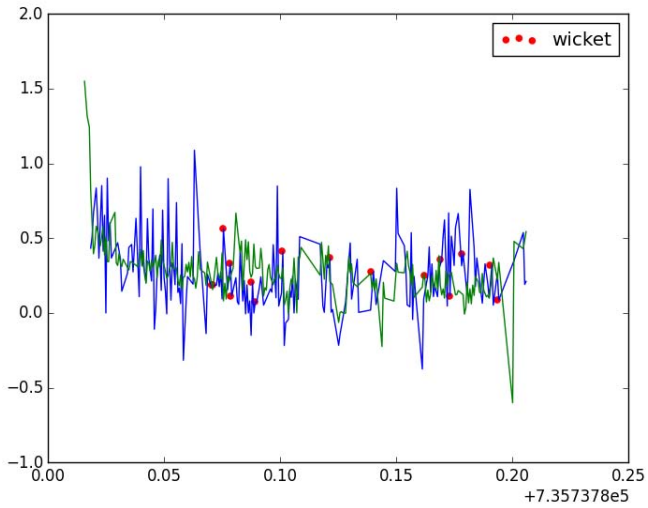

Figure 4.4: Event Detection and Sentiment analysis Result for Qualifier1

5. Conclusion and Summary

The system requires identifying sub-events in match and uses burst in tweet rate to identify it. The system analyzes sentiment using supervised approach like Support Vector Machine, Naive Bayes and Logistic Regression. The systems summarize tweets to find most informative tweet about event, which requires forming clusters and scoring tweets. The system analyzes various queries like finding team fans, sentiment of fan with respect to team, most talked sub-event and etc. This paper concludes with a discussion of potential future work. Firstly, use of paraphrases would improve the detection of low volume events which don't have a concrete vocabulary associated with them. Secondly, use of Out-Of-Vocabulary (OOV) processing is likely to improve the output of a system due to the poor grammar and vocabulary used by Twitter users during live events. Another area of future work would be in machine learning to create a topic-conditioned classifier. Another area of future work would be to identify sentiment with respect to events.

References

[1] Chakrabarti Deepayan and Kunal Punera, "Event Summarization Using Tweets," *ICWSM 11*, pp. 66-73, 2011.

[2] Lin Chen and Chun Lin, "Generating event storylines from microblogs," *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 210-216, 2012.

[3] Nichols Jeffrey, Jalal Mahmud and Clemens Drews, "Summarizing sporting events using twitter," *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pp. 122-128, 2012.

[4] Shen Chao, Fei Liu, Fuliang Weng and Tao Li, "A Participant-based Approach for Event Summarization Using Twitter Streams," *In HLT-NAACL*, pp. 1152-1162, 2013.

[5] Zubiaga Arkaitz, Damiano Spina, Enrique Amigó and Julio Gonzalo, "Towards real-time summarization of scheduled events from twitter streams," *In Proceedings of the 23rd ACM conference on Hypertext and social media*, pp. 319-320, 2012.

[6] Kubo Momoji, Ryohei Sasano, Hiroki Takamura and Minoru Okumura, "Generating live sports updates from twitter by finding good reporters," *In Web Intelligence (WI) and Intelligent Agent Technologies (IAT) 2013 IEEE/WIC/ACM International Joint Conferences on IEEE*, vol. 1, pp. 527-534, 2013.

[7] Corney David, Carlos Martin and Ayse Göker, "Two sides to every story: Subjective event summarization of sports events using Twitter," *In ICMR2014 workshop on Social Multimedia and Storytelling*, pp. 662-672, 2014.

[8] Brendon O'Connor and Balasubramanyan, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series," *Proceedings of the International AAAI Conference on Weblogs and Social Media*, Washington, DC, pp. 511-519, May 2010.

[9] Anjaria Malhar and Ram Mohana, "Influence factor based opinion mining of Twitter data using supervised learning," *In Communication Systems and Networks (COMSNETS)*, 2014 Sixth International Conference on, IEEE, pp. 1-8, 2014.

[10] Turney and Peter D, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," *In Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 417-424, 2002.

[11] Pang Bo and Lillian Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," *In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp. 115-124, 2005.

[12] Baccianella Stefano, Andrea Esuli and Fabrizio Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," *In LREC*, vol. 10, pp. 2200-2204, 2010.

[13] Phillip Stone, "Sentiment lexicon General Inquirer: A Competetive Approach to content Analysis," The MIT Press, 1966.

[14] Bollen Johan, Alberto Pepe and Huina Mao, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," arXiv preprint arXiv: 0911.1583, 2009.

[15] Seol Yong-Soo, Han-Woo Kim and Dong-Joo Kim, "Emotion recognition from textual modality using a situational personalized emotion model," International Journal of Hybrid Information Technology 5, vol. 2, pp. 169-174, 2012.

[16] Hatzivassiloglou Vasileios and Kathleen McKeown, "Predicting the semantic orientation of adjectives," In Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics, pp. 174-181. Association for Computational Linguistics, 1997.

[17] Jiang Long, Mo Yu, Ming Zhou, Xiaohua Liu and Tiejun Zhao, "Target-dependent twitter sentiment classification," In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 151-160, 2011.

[18] Sharifi Beaux, Mark-Anthony Hutton and Jugal Kalita, "Automatic summarization of twitter topics," In

National Workshop on Design and Analysis of Algorithm, Tezpur, India, 2010.

[19] Sakaki Takeshi, Makoto Okazaki and Yutaka Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors," In Proceedings of the 19th international conference on World Wide Web, pp. 851-860, 2010.

[20] Vikrant Hole and Mukta Takalikar, "A Survey on Sentiment Analysis and Summarization for Prediction," In International Journal of Engineering and Computer Science (IJECS), ISSN: 2319-7242, Volume 3, Issue 12, December, 2014, Page No. 9503-9506.