Analysis Of Tweets For Popularity Detection Of Television Media In Business Intelligence

M-Tweets(Media-Tweets)

Dhananjay C. Dandapat
Computer Engineering
Pune University
Pune, India

Sachin C. Chavan
Computer Engineering
Pune University
Pune, India

Narendra P. Chaudhary
Computer Engineering
Pune University
Pune, India

Vaibhav D. Ghare
Computer Engineering
Pune University
Pune, India

***Abstract*—In today's world Social media plays a key role in recording our day to day life responses. Twitter- A social platform serves as a rich source of information for the responses of audiences to any Events broadcasted on television media. Here in this case we propose a system which analyzes particular TV show based on its available #hashtag, extract the tweets classify them as Positive or negative based on their nature, assign scores and display the graphical representation of these scores which serves as a data of greater importance for Business Intelligence. Here we study and implement concept of sentiment analysis to the proposed system, to classify extracted tweets as positive, negative and neutral.**

***Keywords*— Tweets, Business Intelligence, Extraction, Processing, Classification, Clustering, Positive count, Negative count, Graph**

I. Introduction

Now a days social media has been considered as the core area for data mining as it contains the user data in the form of comments, reviews, posts, likes dislikes and also the other platforms like Blogs, Forums, fetch with loads of user generated data. For e.g.: Famous website imdb.com contains all the user reviews to the all movies released as well as all the TV shows.

The data on the social media includes the emotions of the user i.e. how positively or negatively the user is writing his comments or reviews. The positivity and the negativity comprise the important attributes depicting user's mood and emotions.

Here in this case we focus on social platform TWITTER-a popular social platform service were user posts their short messages called tweets (limit 140 characters) for their followers and read the tweets of the ones who they are following. Twitter is very much popular now a days in various areas e.g. the election exit polls are deduced by analyzing trends on twitter during pre-election period. Twitter provide user's access to its API's hence twitter serves as a happy hunting ground for the data miners.

Here the proposed system basically extracts the tweets after the tv show episode has been broadcasted on television, using the #hashtag created by the TV show official twitter account. Further these tweets are preprocessed using data mining techniques, classified into positive, negative and neutral categories based on sentiment classifiers created. Scores are assigned to this data and on the basis of the score the system generates the graph of sentiments (positivity, negativity and neutral counts) this graph can be served as of greater importance to the Particular television network to which the TV show belongs for Business Intelligence.

Here we use the concept of sentiment analysis to classify the tweets. The data posted by user contains his/her sentiments or emotions in the form of words. The sentiments can be positive (words like GOOD, AWESOME), negative (words like BAD, DISASTER) and neutral (words like OK, WELL). The words define the sentiments of the user. Suitable classifiers are created to classify the tweets according to the sentiments in the proposed system.

II. Background And Motivation

A. *Previous Work*

Previous work available on twitter data mining mainly focused on the areas which include Trend Detection and Sentiment Mining. The trend detection was based on sudden activity discovery on general twitter platform The aging theory was applied to detect the trending topics discussed in [1].The system detecting trend by identifying set of word that appear frequently in tweets on twitter platform was discussed in [2].Not much work is available on Twitter data mining in media except the case study discussed in [3]. The study of trends and sentiment analysis on tweets during political debate on TV is discussed in [4], [5]. Also prediction of box office sales is discussed in [6]. No any content regarding twitter data mining in media and applying sentiment analysis to it is available.

B. *Twitter API Overview*

Twitter API provides the developers with tools and techniques to extract data from twitter. Twitter API is basically divided into three parts.

- REST API: It used by developers to access user's status data, timeline and the users profile or information.
- SEARCH API: It gives developers access to search the data on twitter.
- STREAMING API: It gives developers real-time access to huge sampled and filtered data on twitter.

C. *Motivation*

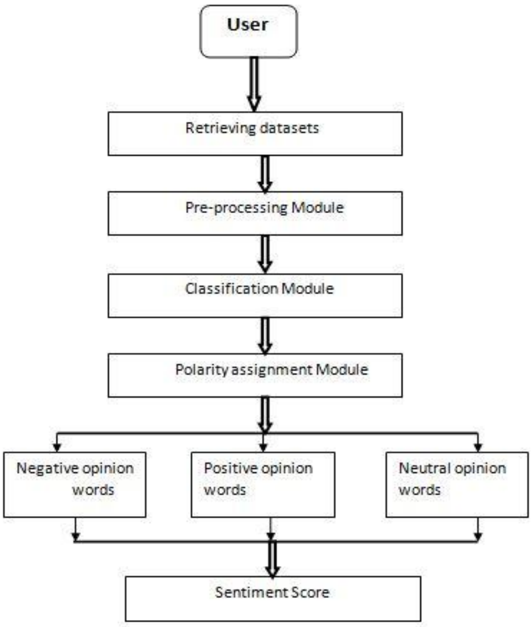
Most of the media networks have their own twitter account. Also each particular TV show has its own separate twitter account. The media network broadcast about upcoming episodes through trailers and teasers through these accounts. Fans mostly react with their view by posting on these accounts. Whenever popular TV show is being aired large amount of real time data i.e. tweets are made by the fans on the handling pages of media network as well as the show page. This real-time data can be collected by the proposed system which can prove of as greater importance to the media network owners as well as show producers to decide further improvements and future changes to the TV show.

III. WORK

The proposed system basically works in 5 steps:

- Tweet Extraction process.
- Data Preprocessing.
- Sentiment Analysis.
- Polarity Assignment
- Graphical representation of Final result.

The figure below explains detail working of proposed system.



A. *The Data(Tweets) Extraction process*

The proposed system uses Twitter4J an unofficial java library for the twitter API with Twitter4J, one can easily integrate the Java application with the Twitter service. The query for extraction can be performed 1 hour after the aired shoe episode ends on TV. Using Twitter4J the system interacts with Streaming API of the Twitter to access the real time tweets from the official page of the TV show also the Search API of the twitter is accessed for firing the search query for e.g. searching the TV show #THEFLASH.

B. *Data Preprocessing*

The extracted tweets are basically written by the users in the way they fill comfortable to express, like use of informal words, adding hyperlinks to the tweets, use of local language etc. The proposed system makes use of Tool like WEKA classifier to preprocess these tweets hence all slang words, hyperlinks, local language etc. are removed from tweets to generate the processed meaningful data also the concept of NLP (Natural Language Processing) is used to design a library to convert the informal words to formal words for e.g. converting gud nght to good night .This processing of data eliminates the unwanted data hence improves the execution time of the system.

C. *Sentiment Analysis*

During this phase the sentiment analysis is done on the processed tweets. Three libraries are defined namely Positive, Negative, Neutral Words library. These libraries are collection of positive, negative and neutral words respectively. These libraries search for the similar words.

From the processed tweets which match with the words present in the libraries. Further counts are assigned to the no of matching words and corresponding clusters of positive negative and neutral words are formed.

D. *Polarity Assignment*

Polarity has been assigned to classified keywords i.e. the scores are assigned which in term is used for plotting the final graph.

E. *Classification Algorithm*

In this step classification algorithm is applied on the keyword from the tweets. This classification algorithm classifies keywords in three classes namely Positive, Negative, Neutral. We use the Quality Threshold (QT) clustering [7] algorithm to cluster words based on similarity measure. Each cluster is a set of words. Further cluster are formed according to the attributes of the word i.e. positive, negative, neutral. To map each tweet to cluster we treat each tweet as query and each cluster as document. We used TF-IDF weighing scheme to calculate distance between tweets and cluster. We then pick the clusters based on highest score [8].

F. *Additional features*

The proposed system aims to work on real time basis hence it provides the graph and score saving feature for offline viewing of the already obtained results. The system also provides with comparison option to compare graph of multiple

TV shows simultaneously thus serving for comparative analysis for Business Intelligence.

IV. Limitations And Future Scopes

A. *Limitations*

The results obtained by using the proposed system are based on the extracted tweets. Only those people who tweet after watching the show are considered in this case. But there exist a vast volume of peoples who don’t use twitter to tweet their views after watching the show. The people tweet are considered to be more techno savvy and adapted to internet hence their sentiments cannot be applied to all general non techno savvy audiences. Also there are multiple social platforms hence only twitter cannot be considered as the representative of all the TV audiences. The case of retweet can lead to analyzing of same tweet no of times it is retweeted hence can lead to non-stable results.

B. *Future Scopes*

The same system can be extended to Facebook and other different social platforms. Facebook can be analyzed since large number of audiences use Facebook as compared to twitter to express their views also the TV shows manage the Facebook, YouTube and also the Instagram pages of their show’s to stay connected with their audiences. This Future Scopes can contribute to the higher accuracy rate of the Proposed System. Beyond hashtag analyzing technique tool can be extended to analyze smiley’s, expressions and images. Advance image processing techniques can be implemented to analyze tweets trends.

Conclusion

Positivity, Negativity are the two important attributes of Human behavior which depict human emotions as well as the mood. This attributes can be used for business intelligence. Here in this case we proposed a system which extract the tweets for a particular TV show, preprocess those tweets. Further the preprocessed data is analyzed for sentiment analysis and the tweets are classified into positive, negative and neutral category, scores are assigned which in term are used to plot graph for TV show progress analysis. Thus this system can serve as of greater importance for Business Intelligence.

Acknowledgment

We express our sincere thanks to Professor Manik Raichurkar (Professor JSPM Narhe Technical Campus) for guiding us throughout the project.

References

[1] M. Cataladi, L. Di Caro, and C. Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. 2010.

[2] M. Mathioudakis and N. Koudas. Twitter monitor: Trend detection over the twitter stream pages 2010.

[3] Business Intelligence from Twitter for the Television on Media: a case study Vignesh T.S, Praveen Kumar. 2010.

[4] N. A. Diakopoulos and D. A. Shamma. Characteri-zing debate performance via aggregated twitter sentiment Volume 2, pages 1195-1198, 2010. A. Shamma, L. Kennedy, and E. F. Churchill. Tweet the debates: Understanding community ann-tation of uncollected sources. Pages 3-10, 2009.

[5] S. Asur and B.A. Huberman. Predicting the future with social media, 2010.

[6] Yooseph S Heyer LJ, Kruglyak S. Exploring Exp-ression Data: Identification and Analysis of Coe-xpressed Genes. Genome Research, Nov1999.

[7] M. De Choudhury, S. Counts, E. Horvitz, "Predicting Postpartum Changes in Emotion and Behavior via Social Media," In Proc. of the SIGCHI Conference on Human Factors in Computing Systems. New York, pp. 3267-3276, May 2013.

[8] Gerard Salton and Michael J. McGill Introduction to Modern Information Retrieval. McGraw-Hills 1983.

[9] 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, 60(11):2169–2188, 2009.