TWITTER TOPIC ANALYSIS USING MULTI-TWEET SEQUENTIAL SUMMARIZATION FOR SENTIMENTAL DATA

Dhuarshine M[1],Archana K[2], Anand Joseph Daniel D[3]

1 *Student, Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India.*
2 *Student, Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India.*
3 *Professor, Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India.*

ABSTRACT

The rise of social media has generated tremendous interest and changes among Internet users today. Data from these social media sites can be used for a number of purposes, like prediction, marketing or sentimental analysis. Twitter is a social networking service on which users post and interact with a message called "TWEETS". The millions of tweets received every year could be subjected to sentiment analysis. But handling such a huge amount of unstructured data is a tedious task to take up. The current Analytics tools and models used that are available in the market, but are not sufficient to manage big data. Therefore, we have utilized Hadoop for intelligent analysis and storage of big data. In this proposed work, we did sentiment analysis on tweets in Hadoop environment.

Keywords*: Streaming data abstraction, Emoticon based opion is analyzed,Non –text feature extraction.*

1. INTRODUCTION

Over the recent years, Twitter has grown from a vague invention to become a mainstream medium for dissemination of messages and the public discussion of news and events. The rapid increases in the amount of Twitter post presents a big obstacle for efficient information acquisition. It is impossible for a user to get an overview of important topics on Twitter by reading all tweets day by day. Additionally, because of information redundancy and the informal writing style, it saves time i.e time consuming to find useful information about a topic from a large number of tweets. The tremendous volume of tweets suggests that summarization is the key to facilitating the requirements of topic exploration, navigation, and search from huge amount of tweets. Specifically, a summary that provides representative information of trending or particular searched topics with no redundancy and well-written sentences would be preferred.

To reduce this problem, many applications have evolved from Twitter (serving as their clients), like “echofon”, “whatthetrend”, which provide services to explain why the term becomes a trending topic or to give a short description of the topic. These systems generally track the topics in Twitter and use existing tweets or encourage users to edit a new tweet to explain the topics. “Whatthetrend” encourages users to edit explanatory tweets about topics. It ranks the submitted explanatory tweets by readers' agreements. These explanatory tweets can be regarded as tiny summaries about the topic, providing a good way to help users understand the topic. However, a short summary can only sketch the topic in a simple way. Some researchers attempt to aggregate several explanations into one long summary using traditional summarization approaches, but it still loses much useful information, such as the change of twitters' focus and the temporal information. A well-generated traditional summary can reflect the overall picture of topic, but performs poorly in summarizing these temporal changes of the crowd's focus in Twitter. Note that the focus of tweets changes much more frequently than that of the traditional mainstream.

LITERATURE SURVEY

LITERATURE SURVEY 1

CONCEPT USED

Metaphor is a pervasive(or common) feature of human language that enables us to conceptualize and communicate abstract concepts using more concrete terminology. In particular, they address the problem of understanding metaphoric language in the context of entailment (or paraphrase) detection. Within a metaphoric context they have performed an in depth experimental analysis to determine which techniques are most effective at interpreting metaphorical text.

LITERATURE SURVEY 2

CONCEPT USED

The primary purpose is to group the tweets by importance or usefulness so that an end user can be presented with a reasonable extract. Summarization is accomplished using a non-parametric Bayesian model applied to Hidden Markov Models and a novel observation model designed to allow ranking based on selected predictive characteristics of individual tweets.

LITERATURE SURVEY 3

CONCEPT USED

Video summarization is a challenging problem because knowing, which part of a video is important requires prior knowledge about its main topic. They observe that a video title is often carefully chosen to be maximally descriptive of its main topic. Hence images related to the particular title can serve as a proxy for important visual concepts of the main topic

LITERATURE SURVEY 4

CONCEPT USED

This approach is inefficient because the spammer might change the phone number or change the content of the text message. Approach is utilizing text classification such as Naive Bayes .k-Nearest Neighbor (kNN).Support Vector Machine (SVM) to recognize pattern of the text messages.

Existing System:

To implement a Trending Topic Analysis System that would analyze the given topic by performing Topic adaptive sentiment classification. Analysis system provides a deep analysis via topic adaptive sentiment classification. Multi tweet sequential summarization, which aims to provide a serial of chronologically.

Disadvantages:

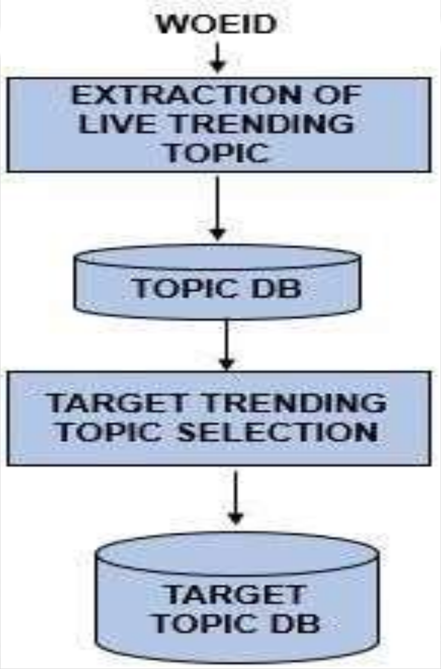
- As the tweet data grows comments as hour by hour it is very typically to analysis sentimental of the users.
- So it is necessary to considered the upcoming present tweet data for the consideration of sentimental analysis

Proposed System:

The proposed system through Foreground Dynamic Topic Modeling. This enables an end user to completely analysis a trending topic to a greater level of detail. Pre-processing phase, the proposed system also handles non-English tweet translation that is essential to prevent discarding of public opinion about the trending topic. It as Stream based approach and Semantic based approach. Redundancy check is performed to remove duplicates from the selected tweets followed by a threshold check to ensure fair mixture of user's opinions**.**

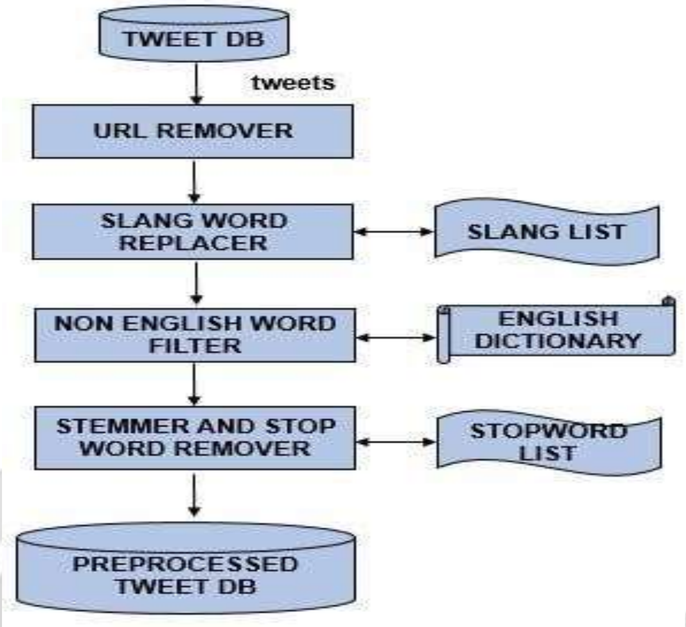
i) Data Extraction

Twitter API will enable extraction of region wise trends. The region‘s “WOEID” which uniquely identifies any region in the world. Trending topics will be stored in the database. While storing the trending topics into the database, they will be tagged with the current during which it trended.


WOEID
EXTRACTION OF LIVE TRENDING TOPIC
TOPIC DB
TARGET TRENDING TOPIC SELECTION
TARGET TOPIC DB

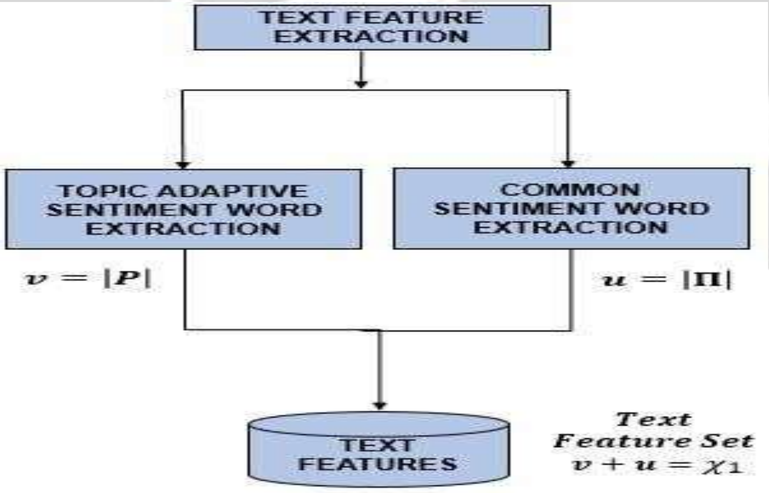
i) Preprocessing Module

Twitter pre-processing involves the following task:
URL removal and link will be removed from the tweet content.
Slang Word Replacement-Slang Words like LOL,OMG will be replaced its appropriate English words i.e, Laugh Out Loud, Oh My God. Non English Words will be filtered with the help of a dictionary. Stemmer and Stop Word removal-English words will be stemmed and only the root words will be retained.



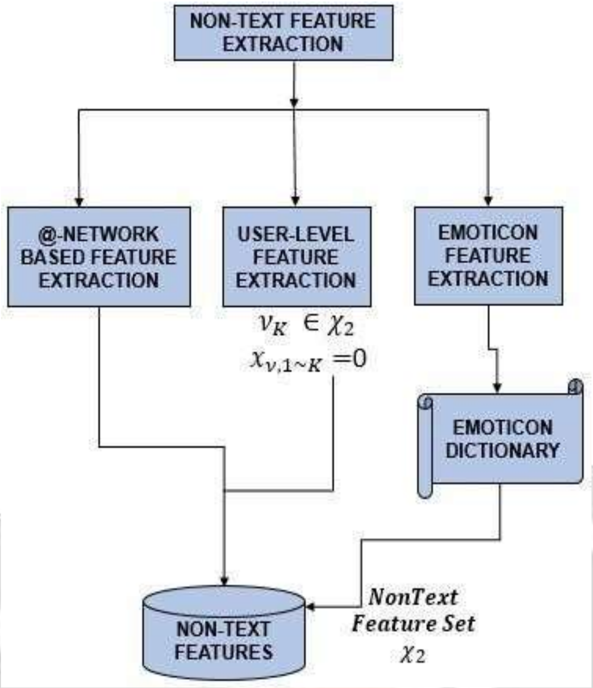
ii) Feature Extraction

Tweets have a special nature unlike normal English sentences like @ symbol used to refer to a user, emotions expressed in the tweets etc. Incorporating such features while performing feature extraction enhances the overall process.



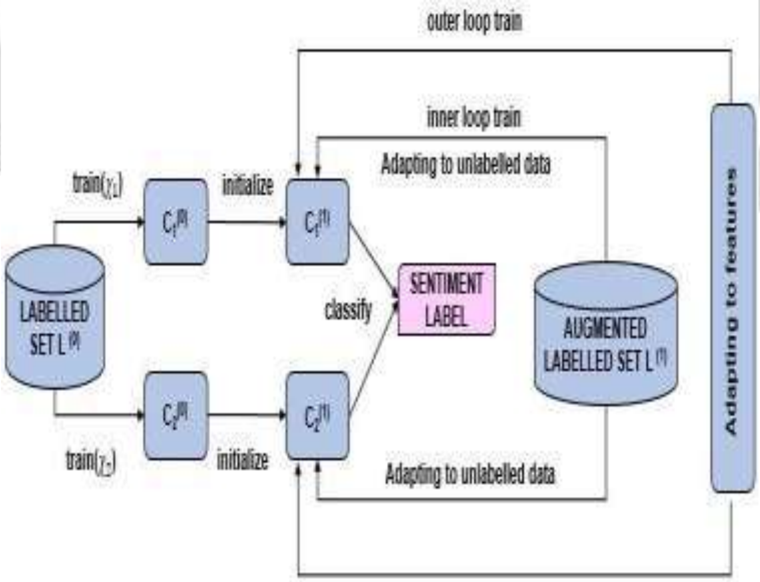
iv) Non- Text Feature Extraction

Text features are a combination of Common Sentiment words and topic sentiment words which will be used to train the classifier. With POS tagging for tweets on a topic and removing the common sentiment words. Proposed framework selects the frequent adjectives, verbs, nouns and adverbs as candidates of topic-adaptive.


NON-TEXT FEATURE EXTRACTION
@-NETWORK BASED FEATURE EXTRACTION
USER-LEVEL FEATURE EXTRACTION
EMOTICON FEATURE EXTRACTION
$v_K \in \chi_2$
$x_{v,1\sim K} = 0$
EMOTICON DICTIONARY
NON-TEXT FEATURES
NonText Feature Set χ_2

iii) Sentiment Classification

This system aims to perform sentiment classification and then using topic wise sentiment labelled data proceeds to the summarization module. Sentimental classification is a topic-sensitive task, i.e., a classifier trained from one topic will perform worse on another. we propose a semi-supervised topic- adaptive sentiment classification model, which starts with a classifier built on common features and mixed labelled data from various topics. Sentiment classification showing the number of correctly classified records and its accuracy.



CHALLENGES FOR TRENDING TOPICANALYSIS

Nature of tweets poses two major challenge namely,

Diversity in Topics

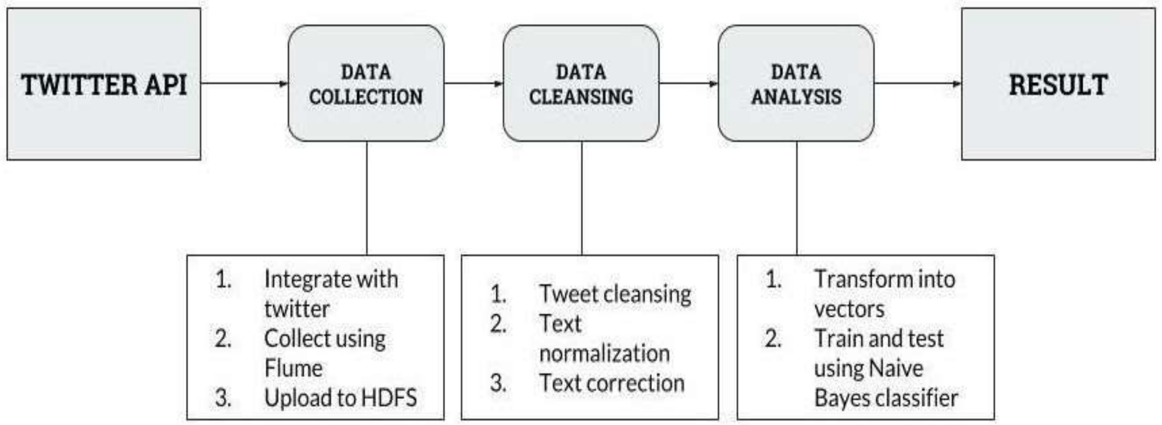
Noise in Tweets

Diversity in Topics

Users discuss range of topics over an increasing time and speed, which poses a major challenge for automatic summarization. Diversity also comes in form of varied form of data to handle.

Noise in Tweets

Tweets poses some unique characteristics, which makes it noisy to handle. Firstly, all tweets are limited to 140 characters. Some tweets are news headlines from the official media; users generate others with various degrees of familiarity with the social media. The resulting tweets can be very different regarding the text quality and word usage. Secondly, Tweets lack structure information, contain various ill-formed sentences and grammatical errors. There are lots of noisy nonstandard tokens, such as abbreviations (“feelin’” for “feeling”), substitutions (“Pr1mr0se” for “Primrose”), emoticons, etc.



SYSTEM ARCHITECTURE

The system architecture is illustrated in the proposed framework analyses a trending topic based on a particular region. Execution of the proposed may be broadly viewed into three categories. Target data collection which involves extracting trending topic and filtering the selected topics for further analysis. Following it is the tweet extraction based on the topic. Then, Pre- processing module cleans and prepares the workable data set. Pre-processed tweets are then fed into the sentiment classifier which will be labelled adaptively. Classifier trained is evaluated as an online process. Sentiment classification process address the issue of generation of conflicting summaries. Finally, the sequential summarization phase in this project handles Topic evolution issue. Latent topics are hidden in the dataset which needs an efficient topic detection models. In this project Stream based sub topic detection models and Semantic based sub topic detection models help in achieving the task of discovering the sub topics within the trending topic. Then, using graph-based approach significant tweets will be selected and topic wise sub summaries are generated in chronological order and presented as extractive sub summaries.

Conclusion

Hash tags, we can provide a simple automated method to what people think. Thus collecting information from networks and analyzing it using Big Data techniques has behind the traditional database approach. Sentiment analysis of Twitter using big data helped us to analyze huge amount of datasets We can further compare the of various providers and judge which one is the best. In order to measure the performance with the existing system we initially implement with Map reduce architecture. In the next phase we will implement with Spark frame with and the performance will be analyzed.

Future Enhancements

They contain 75 million opinions and reviews worldwide. Sentiment analysis helps such websites by converting dissatisfied customers into promoters by analyzing this huge volume of opinions. One more feature we that is worth exploring is whether the information about relative position of word in a tweet has any effect on the performance of the classifier. It can attempt to perform separate sentiment analysis on tweets that only belong to one of these classes (i.e. the training data would not be general but specific to one of these categories) and compare the results we get if we apply general sentiment analysis on it instead.

REFERENCES

[1].Monu Kumar, Dr. AnjuBala, ‘Analyzing Twitter Sentiments through Big Data’ IEEE 2016.

[2].Kim,M.H.Yang,Y.J.Hwang,S.H.Jeon,K.Y.Kim,I.S.Jung,C.H.Choi, W.S.Cho and J.H. Na, ´Customer Preference Analysis Based on SNS data´, 2012 Second International Conference on Cloud and Green Computing ,pp.106-113,2012.

[3].AnisZarrard, AbdulazizAljaloud, Izzat Alsmadi,³The Evaluation of The Public Opinion´,IEEE/ACM 7th International Conference on Utility Cloud Computing,2014.

[4].Ye Wu, Fuji Ren,³Learning Sentimental Influence in Twitter´, International Conference on Future Computer Sciences and Application, 2011

5].M.SaravananM,D.Sundar and S.Kumaresh, ³Probing of Geospatial Stream Data To Report Disorientation´, IEEE Recent Advances in Intelligent Computational Systems(RAICS),2013

[6].BeimingSun,Vincent TY Ng, ³Analyzing Sentimental influence of Posts on Social Networks´, Proceedings of the 2014 IEEE 18th International Conference on Computer Supported Cooperative Work in Design.

[7].Masahiro Ohmura,KohKakusho, Takeshi Okadome, ³Social Mood Extraction from Twitter posts with Document Topic Model.

[8].R.Sanjay,´Big Data and Hadoop with components like Flume,Pig,Hive and Jaql´,2013.