Effective Event Detection and Identification Of Influential Spreaders In Social Media Data Streams

[1] S.Kavitha Bharathi, [2]V.Hariharan
[1] Assistant Professor, [2] PG Student
[1,2] Department of Computer application
[1,2]Kongu Engineering College
Perundurai, Erode-638060
Tamil Nadu, India

Abstract : Micro blogging, a well-liked social media service platform, has become a brand new data channel for users to receive and exchange the foremost up-to-date data on current events. Consequently, it's an important platform for detective work new rising events and for distinctive potent spreaders World Health Organization have the potential to actively publicize data concerning events through micro blogs. However, ancient event notification models need human intervention to detect the quantity of topics to be explored, that considerably reduces the potency and accuracy of event detection. In addition, most existing strategies focus solely on event detection and are unable to spot either potent spreaders or key event-related posts, therefore creating it difficult to trace important events in a timely manner. To address these issues, we tend to propose a Hypertext-Induced Topic Search (HITS) based mostly Topic-Decision technique (TD-HITS), and a Latent Dirichlet Allocation (LDA) based mostly Three-Step model (TS-LDA).TDHITS can automatically detect the number of topics as well as identify associated key posts in a large number of posts. TS-LDA will establish potent spreaders of hot event topics supported each post and user data. The experimental results, employing a Twitter dataset, demonstrate the effectiveness of our projected strategies for each detective work events and distinctive potent spreaders.

I. INTRODUCTION

Over the recent years, Twitter has grownup from a imprecise invention to become a thought medium for dissemination of messages and therefore the discussion of reports and events. The speedy proliferation of Twitter posts presents an enormous obstacle for economic data acquisition. It is impossible for a user to get an overview of important topics on Twitter by reading all tweets every day. In addition, because of information redundancy and the informal writing style, it is time consuming to find useful information about a topic from a huge number of tweets. The tremendous volume of tweets suggests report because the key to facilitating the necessities of topic exploration, navigation, and search from many thousands of tweets. Specifically, a summary that provides representative information of topics with no redundancy and well-written sentences would be preferred. Analysis of topics is truly strengthened by performing a sentiment classification but summarization applications lack this component and as a result produce conflicting summaries. Topics discussed in Twitter are very diverse and unpredictable. Sentiment classifiers always dedicate themselves to a specific domain or topic. Namely, a classifier trained on sentiment data from one topic often performs poorly on test data from another. Most of the recent applications for sentiment analysis deal with a single topic under interest. Another concern regarding sentiment classification is if carried out as a supervised approach it will incur heavy manual effort in annotation.

II. RELATED WORKS

1. G. Mane et al proposed a combined approach of Phrase Reinforcement algorithm along with Word Sense Disambiguation and Textual Entailment techniques for generating one line summary. Phrase Reinforcement algorithm aimed at constructing a graph that helps in identifying the most commonly occurring phrases for a central topic by simply searching for the most weighted set of paths through the graph. This methodology lacked temporal nature of summaries and created coherence issues in the summary generated.

2. D. Wen et al describe accomplished Summarization using a non-parametric Bayesian model applied to Hidden Markov Models. A novel observation model was designed to allow ranking based on selected predictive characteristics of individual tweets. Major focus was to investigate the possibility of using a temporal probabilistic data model known as Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM) to process a stream of tweets pertaining to a single subject and cluster the tweets into groups or rankings based on the value of the individual tweets .But Summaries generated did not project the temporal nature.

3. F. Lui et al proposed a Concept Based optimization framework for topic Summarization using Integer Linear Programming. Target data comprised of original and normalized tweets along with web content relevant to the topic. The focus was not on developing new summarization systems but rather utilizing and integrating diverse text sources to generate summaries that are more informative but there were lack of series of sub events identification to show topic- evolving process.

4. B. O'Conor et al explained Twitter topics by presenting a simple list of messages. An exploratory search application for Twitter called Tweet Motif grouped messages by frequent significant terms and result set is subtopics facilitated navigation and drilldown through a faceted search interface. The topic extraction system was based on syntactic filtering, language modeling, near-duplicate detection, and set cover heuristics. However, the system lacked temporal aspect in summaries generated and topic development was not observed .

5. D. Gao et al proposed generated sequential summaries on a topic using Stream and Semantic based approaches. However, the approach fell short of readability of summaries generated. System failed to capture the opinion expressed in the data thus leading to conflicting summaries .

III. SYSTEM METHODOLOGY

The proposed system "Trending Topic Analyzer" is accomplished through topic adaptive sentiment classification and multi tweet summarization. The proposed system aims at generating chronologically ordered sub summaries, which throws light in understanding the topic evolution of the trending topics. Topic under study is assumed to contained several hidden sub topics, which are revealed using the proposed system through Foreground Dynamic Topic Modeling. This enables an end user to completely analyze a trending topic to a greater level of detail.

Target data collection is performed by extracting the trending topic on a region basis. Topic under study is assumed to contain several hidden sub topics, which are revealed using the proposed system. This enables an end user to completely analyze a trending topic to a greater level of detail. Necessary pre-processing is performed to prepare the target useable data. Along with the 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.

Pre-processed data is later analyzed through topic adaptive sentiment classification to identify the public opinion. Sentiment labeled data will then be processed for sequential summarization. Sub topic detection is achieved using Stream based approach and Semantic based approach. Finally, sub summary candidate selection is executed which gives the highest scored tweets incorporating certain unique features of tweets. Redundancy check is performed to remove duplicates from sthe selected tweets followed by a threshold check to ensure fair mixture of user's opinions.

A. SYSTEM ARCHITECURE

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 labeled adaptively. Classifier trained is evaluated as an online process. Sentiment classification process addresses 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.

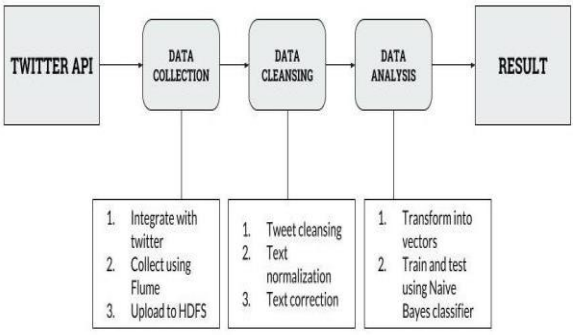


Fig 3.2 System Architecture

B. DATA EXTRACTION

Trending Topic Collection

Trending topics will be extracted using the region‘s “WOEID” which uniquely identifies any region in the world. Twitter API will enable extraction of region wise trends. Extracted trend will be analysed to confirm to the requirement that the topic involves several sub topics hidden in it. Trending topics will be stored in the database. While storing the trending topics into the database, they will be tagged with the month range during which it trended.

This will enable user to pick a topic to analyse based on the month range tagCurrentMonthRange() method mentioned in the following algorithm performs the above mentioned task

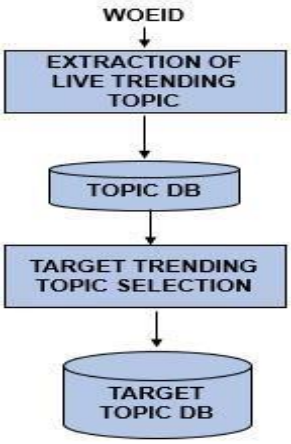


Fig 3.3 Trending topic collection

Topic based Tweet Extraction

Once the trends are extracted, keyword based tweet extraction will be performed using the extracted trend as the keyword Twitter’s JSON response will be parsed and relevant details will be recorded in database.

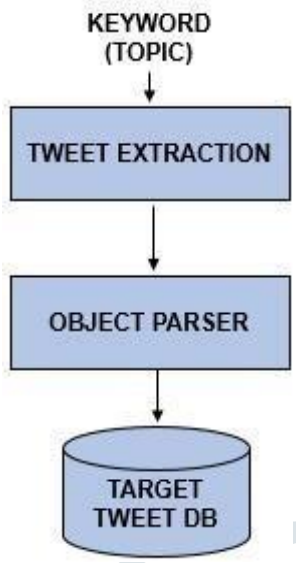


Fig. 3.4 Topic based tweet extraction

C. PRE-PROCESSING

Tweets by nature involve many noisy content and ill-formed words because the users are allowed to tweet only within limited characters. Therefore, cleaning and preparing the meaningful data is essential for any kind of analysis to be performed on the data. Tweet pre- processing involves the following task: URL Removal- URL and links 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 etc. Non-English word filter-

Though tweets could be extracted using specific language, some noisy terms and words corrupt the target data. 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. Stop words in the tweet content will be removed.

Detailed flow of the pre-processing steps is shown in Figure 3.5 along with the algorithm steps described as below.

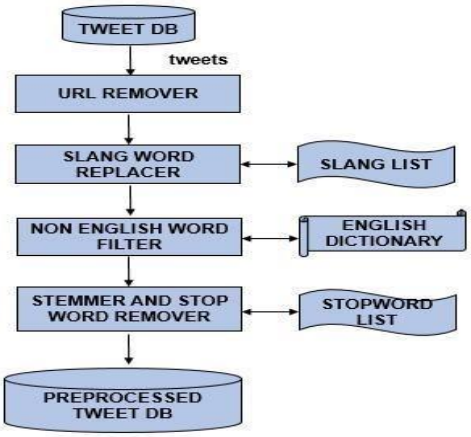


Fig.3.5 Tweet pre-processing

D. FEATURE EXTRACTION

Feature extraction involved in this system creates a significant impact in the performance of sentiment classification task. Tweets have a special nature unlike normal English sentences like @ symbol used to refer to a user, emoticons expressed in the tweets etc. Incorporating such features while performing feature extraction enhances the overall process.

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

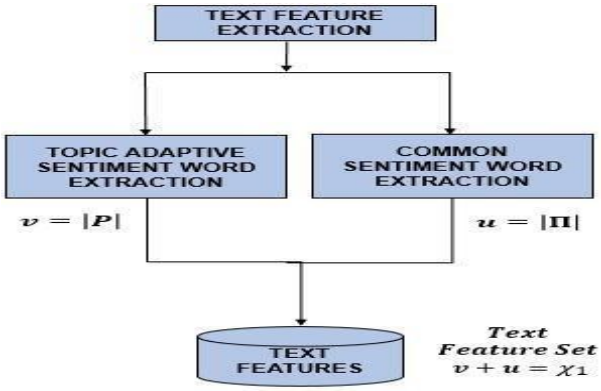
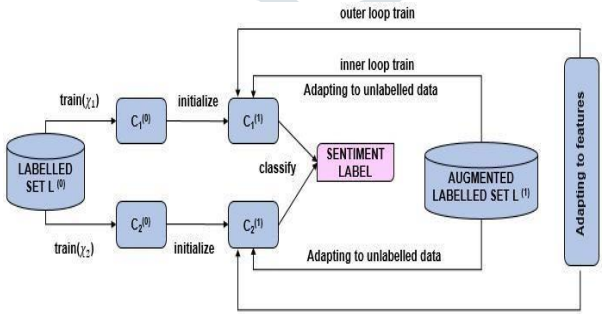


Fig. 3.6 Text feature extraction

E. SENTIMENT CLASSIFICATION

In traditional system, summaries may content conflicting content that reduces the readability And the overall impact of summaries generated. Partially similar to, this system aims to perform sentiment classification and then using topic wise sentiment labeled data proceeds to the summarization module. Sentiment classification may be a topic-sensitive task, i.e., a classifier trained from one topic can perform worse on another. This is particularly a drag for the tweets sentiment analysis. Since the topics in Twitter square measure terribly numerous, it's not possible to coach a universal classifier for all topics. Moreover, compared to product review, Twitter lacks data labeling and a rating mechanism to acquire sentiment labels. The very thin text of tweets conjointly brigs down the performance of a sentiment classifier. classification model, that starts classifier designed on common options and mixed tagged Information from numerous topics .Semi supervised support vector machines S3SVM is one of the most promising candidates to utilize unlabeled data combining with a small amount of labeled ones, since SVM minimizes the structural risk. In addition, collaborative training (co-training) framework is an alternative wrapper, and achieves a good performance, which is often used in the scenarios whose features are easily split into different views. This system adopts the text and non-text features, x1 and x2 as independent views for co-training. In co-training scheme, two classifiers C1 and C2 are trained based on x1 and x2 separately using initial labeled data L. The corresponding feature values are denoted as x and x' respectively for text and non-text feature values. The unlabeled data are selected collaboratively to augment labeled data set L, which is used for the next iteration. In addition, the classifier trained on the combining features using augmented labeled data set L. obtains the final sentiment classification result.

Fig. 3.6 sentiment classification module



IV. CONCLUSION

Social media is a crucial part of the world and it will only keep on building its influence in the future. Analysis of this humongous datasets can improve many networks and industries. We have used one way of analysis called sentiment analysis; there are many ways to improve its accuracy by deploying large datasets considering the emoticons and internationalization.

Many futuristic compressor and filter algorithms will play a major role in bridging the language barrier and in handling noisy content and ill-formed words. Sentiment analysis of Twitter using big data helped us to analyze huge amount of datasets.

V. FUTURE ENHANCEMENTS

Apache Spark may be a quick and general cluster system. It provides high- level genus Apis in Java, Scala, Python and R, and an optimized engine that supports general execution graphs. It conjointly supports an expensive set of higher-level tools together with Spark SQL for SQL and structured processing, MLlib for machine learning, GraphX for graph processing, and Spark Streaming.

In future, Spark tool can be used to improve the performance as Spark is 100 times faster when compared to Map Reduce.

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. Although Pang et al. explored a similar feature and reported negative results, their results were based on reviews that are very different from tweets and they worked on an extremely simple model.

In this analysis focusing on general sentiment analysis. There is potential of work in the field of sentiment analysis with partially known context. For example we noticed that users generally use our website for specific types of keywords which can divided into a couple of distinct classes, namely: politics/politicians, celebrities, products/brands, sports/sportsmen, media/movies/music. So it will plan to perform separate sentiment analysis on tweets that solely belong to 1 of those categories (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. Last but not the least, we can attempt to model human confidence in our system. For example if we have 5 human labellers labeling each tweet, we can plot the tweet in the 2- dimensional objectivity / subjectivity and positivity negativity plane while differentiating between tweets within which all five labels agree, solely four agree, solely three agree or no majority vote is reached. We could develop our custom cost function for coming up with optimized class boundaries such that highest weightage is given to those tweets in which all 5 labels agree and as the number of agreements start decreasing, so do the weights assigned. In this way the effects of human confidence can be visualized in sentiment analysis.

VI. 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.