ISSN 2395-1621

Intelligent Query Answering and Effective Information Access

[#1]Suraj Patil

[1]patil.surajs@gmail.com

Bansilal Ramanth Agarwal Charitable Trust's
Vishwakarma Institute Of Technology
Pune – 411037

ABSTRACT

Twitter and other social networking sites have recently evolved as the main source of data for analysis of a product or a particular thing over the time. These sites have millions of user's everyday producing trillions of data day in and day out. These posts can used for performing sentiment analysis and also for checking what the user are saying about a particular product or a particular feature of that product i.e. positive, negative or neutral. This idea can be tapped for the benefit of extracting opinion on several important hidden characteristics and features about the queried object. In this article we describe an approach that encourages a deeper examination of the contents of the document set retrieved in response to a searcher's query. The approach shifts the focus of inspection and interaction away from potentially uninformative document surrogates (such as titles, sentence fragments and URLs) to actual document content, and uses this content to drive the information seeking process. Current search interfaces assume searchers examine results document-by-document. In contrast our approach extracts features, ranks and presents the contents of the document set. The system will be fetching live data from social n/w sites and generating result as positive, negative and neutral.

Keywords: **Sentiment analysis, Data mining, opinion mining, clustering, ranking documents, Hadoop, Big data**

ARTICLE INFO

Article History

Received: 3rd July 2016

Received in revised form :

3rd July 2016

Accepted: 5th July 2016

Published online :

6th July 2016

I. INTRODUCTION

Today's modern life is totally based on Internet. Now a day's people cannot imagine life without Internet. From last few years people share their views, ideas, information with each other using social networking sites. Users have ability to keep in touch with his/her friends by exchanging different types of information or messages also. Social Media is one of the most significant information exchange technology of the 21st century. People of all ages use social media to share they views and opinions with friends or the wider social web. Social media, such as Twitter or Facebook they can share their ideas / views and opinions of users related to any public topics. Consequently, sentiment analysis of social media content may be of interest to different public sector organisations, especially in the security and law enforcement sector. The results from the above analysis can facilitate government entities and public service organisations to better understand the people they serve and the effect of their actions, as well as to identify potential issues in a timely manner.

The value of systems that help Web searchers find relevant information is becoming increasingly apparent. Our Approach is not only to give a mere dump of information but provide an organized ranked and classified data to highlight the important and diverse aspect of the searched topic. We also provide graphical representation of stats about the topic for easy understanding. The system gives a generic opinion on the topic of interest based on sentiments and importance of each related data. A lot of data/info is generated by social n/w sites; we can use this data to get personalized and reliable data to answer user queries. By using quotes,

© 2015, IERJ All Rights Reserved

status updates, shares etc. of your friends and acquaintances we provide you a different angle for the result(user can now link and associate certain features about the topic of interest with the person's nature whose quote or statement appears in the result). Each #hashTag may have 1000 of comments and new comments are added every minute, in order to handle so many tweets we are using Apache hadoop framework.

II. RELATED WORK

Sentiment analysis has been the field of interest for many researches over the past few years, as the use of OSN is increasing day by day, many people now use OSN to view their opinion about a particular company, product or service. This data can be used for analysis and improving the companies sales.

Dmitry Davidov, Oren Tsur & Ari Rappoport. Provided a supervised sentiment classification framework which is based on data from Twitter.. By utilizing 50 Twitter tags and 15 smileys used as sentiment labels, this framework avoids the need for labor intensive manual annotation, allowing identification and classification of the diverse sentiment types of short texts. They evaluate the contribution of different feature types of sentiment classification and show that their framework successfully identifies sentiment types of untagged sentences. They utilized 50 Twitter tags and 15 smileys as sentiment labels which allow them to build a classifier for dozens of sentiment types for short textual sentences. In their study they use four different feature types (punctuation, words, n-grams and patterns for sentiment classification and evaluate the contribution of each feature type for this task. They showed that their framework successfully identifies sentiment types of the untagged tweets.

Luciano Barbosa provided a 2-step sentiment analysis classification method for Twitter, which first classifies messages as subjective and objective, it further distinguishes the subjective tweets as positive or negative. To better utilize these sources, he verified the potential value of using and combining them, providing an analysis of the provided labels, examine different strategies to combine these sources in order to obtain the best outcome; and, proposed a more robust feature set that captures more abstract representation of tweets, composed by meta-information associated to words and specific characteristics of how tweets are written.

Sascha Narr, Michael Hulfenhaus and Sahin Albayrak provided examined a language-independent sentiment classification approach. They trained a classifier to label the sentiment polarity specifically of tweets. They used a semi-supervised emoticon heuristic to generate labelled training data. For any language, their approach requires only raw tweets of that language for training and no additional adjustments or intervention. They trained classifiers on tweets of 4 different languages: English, German, French and Portuguese. For their evaluation, they collected thousands of human-annotated tweets in these 4 languages using Amazon's Mechanical Turk2.

Preslav Nakov,Zornitsa Kozareva, Alan Ritter, Sara Rosenthal,Sara Rosenthal,Theresa Wilson. Proposed SemEval-2013 Task 2: Sentiment Analysis of Twitter, which included two subtasks: A, an expression-level subtask, A and B, a message level subtask. They used crowdsourcing on Amazon Mechanical Turk to label a large

Twitter training dataset along with additional test sets for Twitter and SMS messages for both subtasks. The primary goal of our SemEval-2013 task 2 has been designed for promoting research that will lead to a better understanding of how sentiment is conveyed through Tweets and SMS messages. Toward that goal, authors created the SemEval Tweet corpus, which contains Tweets (on both training and testing) and SMS messages (for testing only) of sentiment expressions annotated with contextual phrase-level polarity as well as an overall message-level polarity

Anna Jurek, Yaxin Bi, Maurice Mulvenna provided a lexicon based approach for analysing the sentiments of tweets on twitter. They have provided a algorithm that provides the intensity of the sentiments rather than the positive and negative label. They evaluated evidence-based combining function that supports classification process in cases when positive and negative words co-occur in a tweet. They have illustrated a case study of the relation in between sentiment of twitter post related to English defence league

Erik Cambria provided approach for concept level sentiment analysis for automatic analysis of online opinions of various user's using natural language text by machines to go beyond the mere word level sentiment analysis of texts and provide approaches for opinion mining and sentiment analysis that enable's a more efficient passage from textual information to machine process able data node.

III. PROPOSED SYSTEM

In this article we describe and evaluate an approach that encourages a deeper examination of documents at the results interface and blurs inter-document boundaries. We shift the focus of interaction from document surrogates to document content, and rank this content regardless of its source. For this purpose we extract the most common and popular features about the searched topic and rank based on the creation time, likes, shares and owner of the post and presented in a list to the searcher. These are the most potentially useful sentences in the documents, extracted and scored according to factors such as the words they contain (those emphasised by the Web page author, e.g., emboldened terms, and words in the document title or document headings are preferred), and the proportion of query terms they contain. The latter component – scoring by query for terms– are also matched to the synset of the nouns in query this ensures that the sentences extracted are query-relevant. Through presenting the sentences chosen from each document in a feature classified, ranked list, ranked with respect to sentence score and independent of source document, we present a query-

© 2015, IERJ All Rights Reserved

biased overview of the retrieved set's content. In this way, highly relevant content from lower ranking documents, that might not have been viewed, simply because of its resident document's rank position, is made accessible to the searcher.

This project mainly deals with accessing the social networking posts from the users account. The user query is processed to obtain the subject of interest and then all the synsets of the same are obtained. After accessing the account, relevant posts are extracted and processed for removal of stop words, Stemming, Categorizing the post on the basis of association rules and apriori algorithm, applying sentiment analysis on the basis of "POSITIVE" "NEGATIVE" "NEUTRAL" and also ranked according to creation date, shares, likes and owner of post.

IV. ALGORITHM

- Display home page and allow user to login into a social n/w site.
- Open a secure connection and authenticate the user on the social site. Configure the app to use the methods provided by the social site to communicate data.
- Accept the user query.
- Perform basic natural language processing algorithms.
- Get the area of interest from the query by extracting the nouns from query.
- Use WordNet to obtain similar words to expand our search space.
- Calculate the TFIDF score of the data.
- Apply Association Rule to identify interesting relations b/w various features of the topic of interest. Extract diverse important features about the topic of interest.
- I have used sentiment analysis to get the sentiments of each sentence.
- Then I have used cosine similarity to group sentences.
- The Apriori Algorithm is used to obtain the main title for various clusters formed.
- We use a novel ranking algorithm based on time of post, likes and shares.
- We then use R-language to generate charts about the statistics
- Ranking algo used:-
    - Weighted Score=
    - Time + Likes + No. of shares
    - Normalized Score=
    - (nTimes*w1 + nLikes*w2 + nShares*w3)/(w1+w2+w3)
    -

V. EXPERIMENTAL RESULT AND ANALYSIS

The system is expected to give accurate result for analysis of the searched query and sentiments in the form of pie charts and graphs; the system uses two approaches to solve the problem using the normal approach and another using mysql component. The one with mysql component is expected to be more efficient and faster as compared to the normal system in comparison with large amount of data. The system was given 3 types of input file size, small medium and large. The same inputs where passed and processed using normal approach and mysql integrated approach.

The result of normal search engines are compared with the systems searched results by performing experiments on users. Users are asked to acquire knowledge about a particular topic using a precursory search engine and then our system within a given time limit. An average of 31% more knowledge is obtained using our system.

VI. SYSTEM ARCHITECTURE

In the system proposed after logging into the Facebook account, posts are categorised and sentiment analysis is applied on the post, as shown below in fig 1.

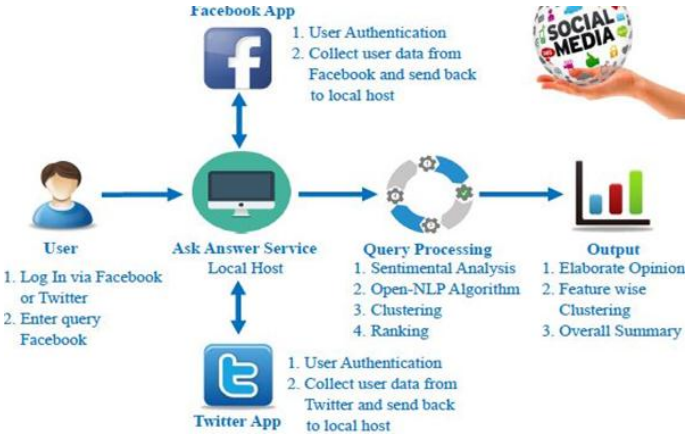

Fig1: System Architecture Diagram

VII. TABLE

Since the data generated and fetched from twitter is in gbs or tbs mysql will surely give the upper hand in execution from the local machines.The below table summarizes the execution time of various file size with and without mysql integration. As twitter/facebook is the largest source of data generation in and around the globe, mysql integration will certainly benefit the analysis and provide effecient and faster results.

File size	With mysql	Without mysql
>5mb	25s	45s
50mb-70mb	62s	45s
100 mb	713s	60s
1gb	998s	257s

Fig 2: Execution time comparison

VIII. SCOPE

The system is very generic and can be used by all the people who want to gain information about a particular thing. The potential candidates could be a buyer of a product who wants to know the key facts and features or

© 2015, IERJ All Rights Reserved

reviews pertaining a particular section of product, it could be a traveler who wants to know better about his destination, it could be a student who wants to get a brief idea about a particular topic and many more.

IX. RESULTS

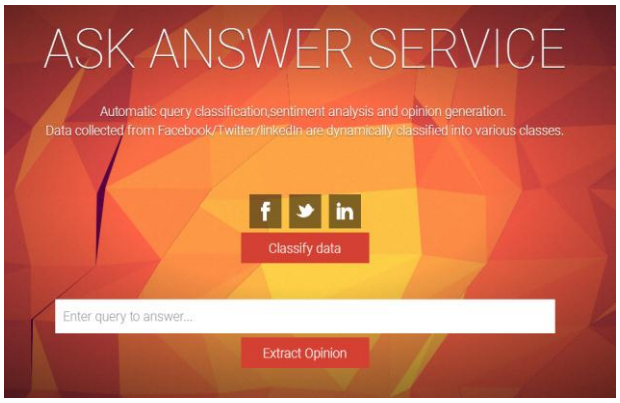

Fig 3: Home Page

fig 4: Result showed in feature wise clusters and in ranked manner

fig 5: Graphical statistics about the result

X. CONCLUSION

Create a generic opinion on the topic of interest based on your social network. Give an insight on various features and aspects about your topic of interest. Try to explore the more subtle and hidden characteristics. Provide ranking to help you focus on most important and useful information. Help you know the real world key aspects and features that matter the most. Provide statistics like weighted scores and confidence on each post.

ACKNOWLEDGEMENT

With deep sense of gratitude we would like to thanks all the people who have lit our path with their kind guidance. We are very grateful to these intellectuals who did their best to help during our project work. It is our proud privilege to express deep sense of gratitude to, **Prof. M. Patwardhan**, H.O.D of Computer Department, for his comments and kind permission to complete this project. We remain indebted to **Prof. Vishal Kaushal**, for their timely suggestion and valuable guidance. The special gratitude goes to staff members, technical staff members, of Computer Tech's. Department for his expensive, excellent and precious guidance in completion of this work. We thanks to all the colleagues for their appreciable help for our working project. With various industry owners or lab technicians to help, it has been our endeavour to throughout our work to cover the entire project work. We also thankful to our parents who providing their wishful support for our project completion successfully. And lastly we thanks to our all friends and the people who are directly or indirectly related to our project work.

REFERENCES

[[1] Dmitry Davidov, Oren Tsur & Ari appoport"Enhanced Sentiment Learning Using Twitter Hashtags and Smileys," Coling 2010: Poster Volume, pages 241– 249,Beijing, August 2010.

[2] Luciano Barbosa, "Robust Sentiment Detection on Twitter from Biased and Noisy Data"Coling 2010: Poster Volume, pages 36–44,Beijing, August 2010

[3] Sascha Narr, Michael H¨ulfenhaus and Sahin Albayrak "Language-Independent Twitter Sentiment Analysis"

[4] Preslav Nakov,Zornitsa Kozareva, Alan Ritter, Sara Rosenthal,Sara Rosenthal,Theresa Wilson." SemEval-2013 Task 2:Sentiment Analysis in Twitter" Preslav Nakov,Zornitsa Kozareva, Alan Ritter, Sara Rosenthal,Sara Rosenthal,Theresa Wilson." SemEval-2013 Task 2:Sentiment Analysis in Twitter"

[5] Anna Jurek, Yaxin Bi, Maurice Mulvenna" Twitter Sentiment Analysis for Security-Related Information Gathering" 2014 IEEE Joint Intelligence and Security Informatics Conference pages 48-55

[6] Erik Cambria, " Sentiment Knowledge Discovery in Twitter Streaming Data,"

© 2015, IERJ All Rights Reserved