ISSN:0975-9646

Mining Social Network Data for Cyber Physical System

Manjushree Gokhale, Bhushan Barde, Ajinkya Bhuse, Sonali Kaklij

Department of Computer Engineering, Savitribai Phule pune university
Gokhale Education Society's R.H.Sapat College of Engineering, Management Research,Nashik-422005

***Abstract*— A major benefit of social media is that you can see the good and bad things people say about your brand or any communication that may trigger terrorism. The bigger your company gets, however, the harder it becomes to keep a handle on how everyone feels about your brand. For large companies with thousands of daily mentions on social media, news sites and blogs, it's extremely difficult to do manually. That's where contents analysis using data mining software comes in. It monitors and evaluates your online mentions to show you how the whole Web is reacting to your news in real time. Our Project statement thus defines to detect whether the given data is sarcastic or not. In the process, a probability is to be calculated that denotes how sarcastic the given tweet is.**

***Keywords*—Cyber Criminal network, Semantic analysis, Naive Bayes Algorithm, .**

I. INTRODUCTION

As sentimental analysis has improved in the last few decade's so have its applications. Sentimental analysis is now being used from specific product marketing to anti social behaviour recognition. The advances in facebook ,twitter , youtube and other micro blogging and social networking sites have not only contributed change to the social sites but have fundamentally changed the way we use these sites and how we share our feelings, our views with the wider audience.

Businesses and organizations have always been concerned about how they are perceived by the public. This concern results from a variety of motivations, including marketing and public relations. Before the era of Internet, the only way for an organization to track its reputation in the media was to hire someone for the specific task of reading newspapers and manually compiling lists of positive, negative and neutral references to the organization. Alternatively, it could undertake expensive surveys of uncertain validity.

Today, many newspapers are published online. Some of them publish dedicated online editions, while others publish the pages of their print edition in PDF or similar formats. In addition to newspapers, there are a wide range of opinionated articles posted online in blogs and other social media. This opens up the possibility of automatically detecting positive or negative mentions of an organization in articles published online, thereby dramatically reducing the effort required to collect this type of information. To this end, organizations are becoming increasingly interested in acquiring fine-grained sentiment analysis from news articles. Fine-grained sentiment analysis is an extremely challenging problem because of the variety of ways in which emotions and opinions can be expressed. News articles present an even greater challenge, as they usually avoid overt indicators of emotion or attitudes. However, despite their apparent neutrality, news articles can still bear a polarity if they describe events that are objectively positive or negative. Many techniques used for sentiment analysis involve naïve approaches based on spotting certain keywords (which reveal the author or speaker's emotions) such as those present in the AFINN word list. Other methods involve combining keywords and other features with machine learning algorithms. However, relying on superficial lexical features such as word choice and syntactical features can only take one so far towards understanding the sentiment behind a piece of text. This is because these approaches do not involve attempting to understand the text itself. The current state of the art is still very far from true natural language understanding. However, researchers are making strides forward in this direction through the usage of common-sense knowledge bases. Common-sense knowledge is obvious information that is usually left unstated and is especially important for analysing news articles, because these generally try to adopt a neutral stance. Polarity must therefore be deduced using affective common-sense knowledge. Combining common-sense knowledge with sentiment analysis can be through semantic computing, which allows emotions to be associated with common-sense concepts. We shall therefore explore the application of semantic computing to fine-grained sentiment analysis in news articles. This project presents an opinion-mining engine we have built which performs fine-grained sentiment analysis to classify sentences as positive, negative or neutral.

II. LITERATURE SURVEY

Negation in sentences may not necessarily make the thought conveyed negative. Negativity not always means use of linguistically negative words, as the message might be negative but 'negative' words like 'not' are not used. Given a large number of user reviews on a topic (IMDB user reviews and ratings was used here) the probability was calculated as:

$P(c|w)$ Where w is the given, and lies in a review with a rating c. Rating is given along with the user review (here it was on scale of 10) The above probability gave a 'polarity' for each word, that gave an inclination to what feelings the user might be conveying in the review. Words like 'no' and 'not' add greatly to the negativity of sentences. Problem arises with some words where different users use them in different contextual meanings. A negative response to a

question, statement or proposal is not necessarily a 'reject'. If the previous statement is phrased in the negative, a 'no' could be an agreement. Words (or phrases) were categorize as emphasizing thoughts or just attenuating towards a thought. Example: "I won't eat spinach" shows some restriction on eating, but "I won't eat any spinach" is stronger, and "I won't eat any spinach at all" is even stronger. Repetitive use of words adds extra emphasis to the meaning of words. Questions about what actually happened in language use have been much less central, though such questions are fundamental to understanding human talk exchanges.

Businesses and organizations have always been concerned about how they are perceived by the public. This concern results from a variety of motivations, including marketing and public relations. Before the era of Internet, the only way for an organization to track its reputation in the media was to hire someone for the specific task of reading newspapers and manually compiling lists of positive, negative and neutral references to the organization. Alternatively, it could undertake expensive surveys of uncertain validity.

Today, many newspapers are published online. Some of them publish dedicated online editions, while others publish the pages of their print edition in PDF or similar formats. In addition to newspapers, there are a wide range of opinionated articles posted online in blogs and other social media. This opens up the possibility of automatically detecting positive or negative mentions of an organization in articles published online, thereby dramatically reducing the effort required to collect this type of information. To this end, organizations are becoming increasingly interested in acquiring fine-grained sentiment analysis from news articles. Fine-grained sentiment analysis is an extremely challenging problem because of the variety of ways in which emotions and opinions can be expressed. News articles present an even greater challenge, as they usually avoid overt indicators of emotion or attitudes. However, despite their apparent neutrality, news articles can still bear a polarity if they describe events that are objectively positive or negative. Many techniques used for sentiment analysis involve naïve approaches based on spotting certain keywords (which reveal the author or speaker's emotions) such as those present in the AFINN word list.

III. PROPOSED SYSTEM

Sentiment analysis classifies the opinions into positive and negative categories. We focus on the technique to detect the topic related to the positive and negative opinions. Knowing the reasons behind classifying the sentiment provides better perception. These reasons are called as sentiment topics associated with the sentiment. The proposed method collects web content and extracts snippets from them. Snippets are keywords like terror, attack any brand names. Then a sentiment score is calculated for each snippet based on which they are classified into different categories to create sentiment taxonomy. Topics related to each category are identified. Point wise mutual information and mutual support are used to find words for a particular topic, to evaluate the importance of a word in a category. Then, the word with highest frequent value and highest point wise mutual information value is chosen as the topic. We propose an approach that is called proximity based sentiment analysis. We proposed a method which considers the negation scope and strength of a word while classifying whether a word has positive or negative effect on the sentence. The proposed approach uses two algorithms; the first one is used to calculate sentence score for each word. In the second algorithm, the sentence score is calculated using the word sense and word score with respect to each negative keyword. If the calculated sentence score is less than zero, then it is assigned to a negative class.

Sentiment analysis classifies the opinions into positive and negative categories. We focus on the technique to detect the topic related to the positive and negative opinions. Knowing the reasons behind classifying the sentiment provides better perception.

IV. ARCHITECTURE

For our news-focused opinion-mining engine, we chose to apply semantic computing, for its ability to leverage on commonsense knowledge and its wide scope for future research.

Semantic computing is an emerging multidisciplinary field pioneered by Cambria and Hussain whose general objective is to enable computers to understand human emotions. In order to achieve this goal, computers require both conceptual information about our world (which can be obtained from common-sense knowledge bases) and the relative affective information associated to it. The system is divided into major phases a) Parsing Phase Pre-processing b) Feel Phase i.e. Sentiment detection.

Four lexicons are used for sentiment analysis (number of terms in the lexicons in brackets): "positive tone" (332), "negative tone" (630), "strength of sentiment" (59), "negations" (45). These lexicons have been created manually by the inspection of thousands of tweets, and continue to be expanded on a regular basis. Note that the same term can appear in different lexicons (if rarely in practice). For example, the term hate appears in the lexicon for negative tone and in the lexicon for strong sentiments. Each term in a lexicon is accompanied by a heuristics and a decision rule.

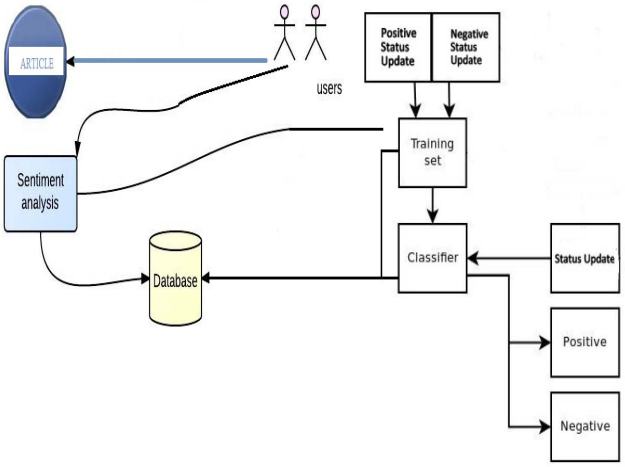

fig 1. Architechture of proposed system

A major problem faced during the task of sentiment classification is that of handling negations. Since we are using each word as feature, the word "good" in the phrase "not good" will be contributing to positive sentiment rather that negative sentiment as the presence of "not" before it is not taken into account. To solve this problem we devised a simple algorithm for handling negations using state variables and bootstrapping. We built on the idea of using an alternate representation of negated. Our algorithm uses a state variable to store the negation state. It transforms a word followed by a not or n't into "not_" + word. Whenever the negation state variable is set, the words read are treated as "not_" + word. The state variable is reset when a punctuation mark is encountered or when there is double negation.

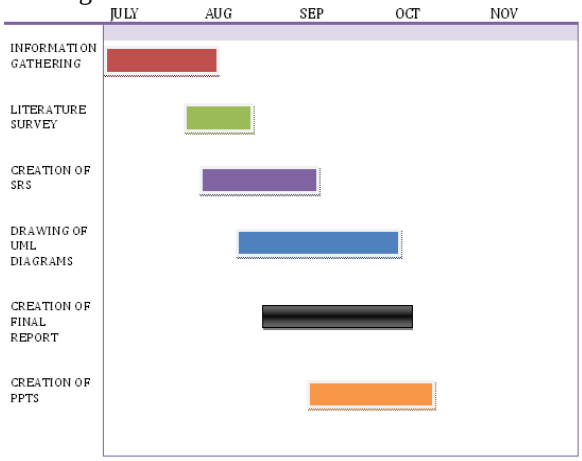

Fig 2. Detailed plan of system implementation

V. NAIVE BAYES ALGORITHM

The Naïve Bayes method for classification is often used in text classification due to its speed and simplicity. It makes the assumption that words (or k-grams) are generated independently of word position.

The classifier then returns the class with the highest probability given the document. In practice, the log probability is estimated, A Naive bayes classifier is a simple probabilistic model based on the Bayes rule along with a strong independence assumption. The Naïve Bayes model involves a simplifying conditional independence assumption. That is given a class (positive or negative), the words are conditionally independent of each other. This assumption does not affect the accuracy in text classification by much but makes really fast classification algorithms applicable for the problem. The frequency counts of the words are stored in hash tables during the training phase independence If we use the simplifying conditional assumption, that given a class (positive or negative), the words are conditionally independent of each other. Due to this simplifying assumption the model is termed as "naïve".

VI. CONCLUSIONS

Thus, we studied about how the cyber criminal effect our society and how we can stop them with the latest technologies, also the system we are going to develop will definitely help us to secure more social media from the effect of cyber criminals.

The proposed computational algorithm can effectively extract semantically rich representations of latent concepts describing transactional and collaborative relationships among cybercriminals based on publicly accessible messages posted to online social media.

ACKNOWLEDGMENT

The work explained in this paper was guided by our respective project guide Mrs.A.S.Vaidya. With all her help it was easy to understand the concept regarding data mining, cyber security.

It helped a lot for understanding and studying different algorithms work on data mining and cyber security domain. We had a very good guidance of our teachers to go through different software and technologies. Understanding new things and using our knowledge we are developing this project considering in mind that definitely this project will help us to reduce such hazardous cyber criminal attacks and to keep surveillance on cyber criminal.

REFERENCES

[1] Raymond Y.K. Lau, Yunqing Xia, Yunming Ye , "A Probabilistic Generative Model for Mining Cybercriminal Networks from Online Social Media" Digital Object Identifier 10.1109/MCI.2013.2291689 Date of publication: 14 January 2014

[2] H. Abbass, A. Bender, S. Gaidow, and P. Whitbread, "Computational red teaming: Past, present and future," IEEE Comput. Intell. Mag., vol. 6, no. 1, pp. 30–42, 2011.

[3] S. Bao, R. Li, Y. Yu, and Y. Cao, "Competitor mining with the web," IEEE Trans. Knowl. Data Eng., vol. 20, no. 10, pp. 1297–1310, 2008.

[4] D. M. Blei, A . Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003.