Analysis of Public Sentiments using Annotation Technique

[1]Syed Shoeb Syed Razique, [2]Bhakti Ahirwadkar, [3]Supriya Kinariwala

[1]PG Student, [2]Associate Professor, [3]Assistant Professor
Department of Computer Science and Engineering,
Marathwada Institute of Technology, Aurangabad, India 431005

***Abstract*: Many users share their comments on Twitter that makes it a usable place for analyzing and interpreting public sentiment. This monitoring and analysis will provide important information for higher efficiency processes across multiple domains. In the proposed system, the model tends to take a step towards interpreting the sensory variations of tweets. We tend to discover that the subject has different feelings in a unit time period. The confidence variance is related to the true reason for the variations. To support this observation, LDA algorithm is used to refine topics and highlight long titles. These highlights will provide an interpretation of the potential trust model. We tend to select the most important tweets for the subject, and to develop another mechanical model, called hybrid model using Triple Relation Extraction (HTRE). To improve the readability of the search, most representative tweets were used for the leading subject and rank topics based on their "popularity" for the transition. The results show that our method can find the foreground and logical order efficiently. The proposed format can be used for other tasks, such as finding the difference between different twitter datasets.**

***Keywords:* Sentiment Analysis, Opinion mining, Twitter, Latent Dirichlet Allocation, Triple Relation Extraction, Gibbs Sampling, Emerging events.**

I. INTRODUCTION

The explosive growth of user-generated messages, Twitter has become a social site where millions of users can exchange their opinions. The sentiment analysis on Twitter data has provided an economical and effective way to expose public opinion in a timely manner, which is essential for decision-making in various areas. For example, a company may study public opinion in tweets for user feedback on its products; while a politician can adjust his position regarding the change of public sentiment. There have been a large number of research studies and industrial applications in the field of public opinion tracking and modeling. Previous research like O'Connor et al. [1] focused on monitoring public opinion on Twitter and examined its correlation with consumer confidence and polls on approval of presidential positions. Similar studies have been conducted to study public opinion on stock markets [4] and oil price indices [3]. They reported that events in real life have indeed a significant and immediate effect on public opinion on Twitter. However, these studies conducted further analysis to exploit useful information on the significant change in sentiment, called variation in public opinion.

A useful analysis is to find the possible reasons for varying feelings, which can provide important information about decision-making. For example, if the negative sentiment towards Person increases significantly, the organization related to that person may be eager to know why people have changed their minds and are reacting accordingly to reverse this trend. Another example is that, if public opinion changes considerably on certain products, the companies concerned may want to know why their products receive such returns. It is usually difficult to find the exact causes of the variations of feeling because they can involve complicated internal and external factors. We observed that new topics discussed during the variation period could be strongly related to the true reasons for the variations. When people express their opinions, they often mention reasons (for example, specific events or topics) that support their current point of view. In this work, we consider these emerging events/topics as possible reasons. Emerging mining events/topics are difficult: (1) the collection of tweets during the variation period could be very noisy, covering irrelevant topics of "interest" that have been discussed for a long time and have not helped to modify public opinion. How to filter these background topics is a problem that we must solve. The techniques of text grouping and synthesis [8], [11] are not suitable for this task because they allow discovering all the subjects of a collection of texts. (2) Events and topics related to the variation of opinion are difficult to represent. Aim to implement a classification algorithm for text in positive, negative or neutral, extract the meaning of a text or a tweet using Natural Language Processing, to determine the attitude of the mass in various objectives vis-à-vis the subject of interest. Improve the accuracy of the analysis.

Keywords generated by topic modeling [2] can describe the underlying events to a certain extent. But they are not as intuitive as natural language sentences. (3) The reasons can be complicated and involve a number of events. These events may not be as important. Therefore, the mined events must be classified according to their contributions. In this proposed system, we analyze the variations of public opinion on Twitter and we explore the possible causes of these variations. To keep up with public opinion, we are using Stanford CoreNLP sentiment analysis tools [26] to get information about feelings about interested targets in each tweet. Based on the sentiment tag obtained for each tweet, we can track public opinion about the corresponding target using descriptive statistics (e.g. percent feeling). On the tracking curves, significant feeling variations can be detected with a predefined threshold (for example, the percentage of negative tweets increases by more than 50 %).

Latent Dirichlet allocation algorithm used to analyze tweets in significant periods of variation and to deduce the possible reasons for the variations. At first part of a proposed system, to filter background topics and extract foreground subjects from tweets during the variation period, using an auxiliary set of tweets generated. By removing the interference of long-time historical subjects, LDA can meet the first challenge. To handle the last two challenges, in a second part called Hybrid Triple Relation Extraction (HTRE). HTRE first extracts representative tweets from leading subjects (obtained from LDA) as candidates for reason. Then, it will associate each tweet remaining in the variation period with a candidate with a reason and rank the candidates by the number of tweets associated. Experimental results on real Twitter data show that our method efficiently extracts the desired information behind the variation of public opinion. In summary, the main contributions of this paper are (1) to analyze and interpret variations in public opinion on microblogging services; (2) solve the problem of mining reason. The proposed system is general: they can be applied to other tasks such as searching for subject difference between two sets of documents.

We perform a linguistic analysis of the collected corpus and explain the phenomena discovered. Using the corpus, we construct a feelings classifier, able to determine the positive, negative and neutral feelings for a document. Experimental evaluations show that our proposed system is effective and works better than previously proposed methods [10].

II. RELATED WORK

A semantic-based friend recommendation system for social networks [5] presents the design and implementation of Friendbook, a semantic recommendation system based on friends for social networks. Unlike friend referral mechanisms using social graphs in existing social networking services, Friendbook has extracted lifestyles from user-centric data collected from sensors on the smartphone and recommended to potential users to share similar lifestyles. Implemented Friendbook on Android-based smartphones and evaluated its performance on small-scale experiments and large-scale simulations. The results show that the recommendations were accurate.

According to the Opinion Extraction and User Opinion Summary Survey: on the Web a large amount of user-generated data is present on the Web in the form of blogs [24], reviews [20], comments, event [2], [15], news [21] and so on. Public opinion (sentiment analysis) [14] is a process of seeking user opinion from user-generated content. The opinion summary is useful for analyzing feedback, making professional decisions and referral systems. In recent years, the mining industry is one of the most popular topics in text extraction and automatic language processing. This paper presents the methods of opinion extraction, classification and synthesis. This paper also explains different approaches, methods, and techniques used in the process of extracting and synthesizing opinions, as well as a comparative study of these different methods [25]. There have been some studies on predicting the movie sales [24] and predicting election [13] result based on public opinion on social site. Also in the interpretation of public sentiment on the social website and relation between real lives events occurred, [6], [7], and [15]. The Twitter platform [10] is useful for following the feelings of the public. Knowing the views and reasons of users at different times is an important study to make certain decisions. The categorization of positive and negative opinions is a process of feeling analysis. It is very useful for mining tweets and summarize with the help of target name, keyword [12], people to find a sentiment polarity about the person, the product etc.

In this proposed system, interpreting public opinion and reasons behind sentiment variation by two methods. When the first method called LDA can remove background topics and select prominent topics among the tweets with the help of heuristics rule and detect non relevance topic and remove them. The second template HTRE [19], which extracts the most representative tweets from LDA algorithm. The candidate is a reason for the interpretation of public feelings. As a corpus for the analysis of feelings and the exploration of opinions on Twitter: the microblogging has become today a communication tool very popular with Internet users. Some research work has been devoted to this topic such as tracking the sentiment variation and finds the reason [9], [10]. In our paper, focus on using Twitter, the most popular microblogging platform, for sentiment analysis and showing how to automatically collect corpus for sentiment analysis and opinion extraction.

III. PROPOSED ARCHITECTURE

In the proposed system, we are using Latent Dirichlet Allocation (LDA) and hybrid triple relation extraction (HTRE). The LDA template can filter background topics and then extract leading topics to reveal possible reasons. To give a more intuitive representation, the HTRE model can classify a set of reasons expressed in natural language to provide phrase level reasons. Another major problem is mining. The volume of opinions is composed of both foreground and background reasons, which is the main difficulty in resolving differences. To improve the readability of mined reasons, we select the most representative tweets for leading topics and develop another model called Hybrid Triple Relation Extraction (HTRE), to rank them according to their popularity in the variation period. Experimental results show that our methods can effectively find prominent subjects and rank the candidates of the reason.

i. Tweets Extraction Process

Extract twitter tweets based on the keywords of the query. In this module based on the Twitter access key and the consumer key, we will extract the tweets based on the keyword of the query. Twitter messages are taken as input, Twitter messages are very informal, these messages are filtered using techniques such as URL filtering, translation of slang words, filtering non-English tweets, deleting of empty words. Preprocessing Tweet is done by removing unnecessary words, removing hyperlinks, removing special characters and filtering data to identify the value of feeling.In this module, we are using Stanford CoreNLP tool [26] for sentiment

analysis process. To retrieve the tweets linked to the target, we can browse the entire dataset and extract all the tweets containing the keywords of the target. Compared to the usual datasets containing tweets, they are usually less formal and written repeatedly ad hoc. The sentiment analysis tools applied to raw tweets repeatedly achieve very poor performance in most cases. Therefore, pre-processing techniques on tweets are necessary to obtain satisfactory results on the analysis of feelings. Stanford's NLP is used for this purpose and the messages are labeled as positive or negative or neutral.

ii. Tracking Sentiment Variation

Once sentiment tags are obtained for all tweets retrieved for a target, the proposed system can track sentiment change using various descriptive statistics. Here, the percentage of positive or negative tweets among all extracted tweets is adopted as an indicator of the tracking of the change in sentiment over time. On the basis of these descriptive statistics, one can find variations of feeling by using different heuristics (example: the percentage of positive / negative tweets increases by more than 50%). we used Stanford's NLP framework that works with PennTree and uses the PTB tokenizer. Initially, the Penn Treebank English style token was extended to support other languages and noisy Web style text, and then implemented as a deterministic finite automaton.

Fast and efficient - 1,000,000 chips per second are supposed to be processed by the structure. The sentence division is a deterministic consequence of tokenization: the sentence ends when an end-of-sentence character (.,!,?) Is found and is not grouped with other tokens. The annotator pipeline is used, once the text needs to be processed, a pipeline provides more control and functionality.

```
Properties props = new Properties();

props.setProperty("annotators", "tokenize");

StanfordCoreNLP pipeline = new StanfordCoreNLP(props);

Annotation annotation = new Annotation("This is a sentence to be tokenized");

pipeline.annotate(annotation);
```

The Properties class = persistent set of properties that can be saved to or loaded from a stream. The property list stores key and values as Strings. The Stanford CoreNLP class is designed to apply multiple Annotators (or functions) to a annotation the annotate() method.
Annotation = a representation of text, either raw or processed.

CoreLabel = Map from keys to values. It provides convenient methods to access tags, lemmas, etc.
TokensAnnotation = CoreMap key for getting the key for getting the tokens contained by the Annotation.

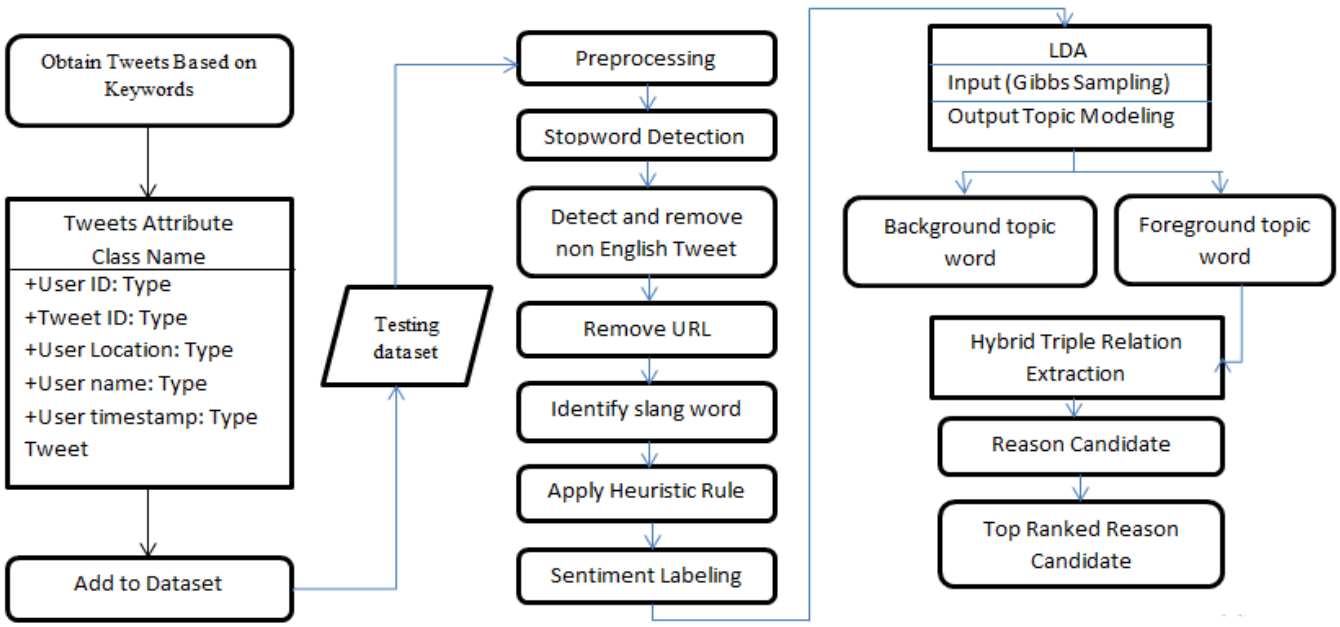

Figure 1.0: Proposed Architecture

iii. Latent Dirichlet Allocation Algorithm

The Latent Dirichlet Allocation (LDA) can filter background topics with the help of raw aspects and extract foreground subjects from tweets during the variation period. The LDA algorithm is used for this purpose. HTRE first extracts representative tweets from leading subjects (obtained from LDA) as candidates for reason. Then, he will associate each remaining tweet in the variation period

with a candidate with a motive and rank the candidates with reason by the number of tweets associated with them. The HTRE algorithm is used for this purpose.

iv. **Steps of LDA Algorithm**

1. Process through each tweet and randomly assign each tweet word to one of the K topics.

2. Note that this random assignment already gives you both topic representations of all tweets and word distributions of all topics.

3. Improve them for each tweet.

4. For each word w in d and for each subject t, calculate two things:

• p (subject t | tweet d) = the proportion of words in the tweet d currently assigned to the subject t,

• p (word w | heading t) = the proportion of assignments to the subject t on all tweets from this word w.

5. Reassign w a new subject, where you choose subject t with probability p (subject t | tweet d) * p (word w | topic t.

6. In other words, in this step, we assume that all subject assignments, except for the current word in question are correct, then we update the assignment of the current word using our template of how generation of tweets.

7. After repeating the previous step a lot of times, we will eventually reach a relatively stable state where your assignments are pretty good. Use these assignments to estimate the mix of topics in each tweet (counting the proportion of words assigned to each topic in this tweet) and the words associated with each topic (counting the proportion of words attributed to each topic).

v. **Steps for generative process of HTRE**

We automatically select the reason candidates by finding the most relevant tweets for each of the top-notch topics learned from LDA [2], using the following metric [10]:

$$\text{Relevance}(t, k_f) = \sum_{i=t} \emptyset_f^{k_f, i} \qquad (1)$$

where $\emptyset_f^{k_f}$ is the distribution of the word for the foreground subject k_f and i is the index of each non-repetitive word in the tweet t.

1. Get top-notch topics and find the relevance of tweets using the formula below:

2. For each tweet, find the word distribution and look for the relevance of the words.

3. Extract the most relevant tweets.

4. View his account and tweet.

The experiment was carried out on an i3 processor using a Java platform on 4 GB of RAM in primary memory. The tweets were collected using the authentication key and the consumer key with the Twitter API via a Twitter account.

The Specific keyword [12] was used to collect tweets related to a specific topic and these tweets were then pre-processed to clean the noise and other degenerative terms in order to get specific feelings about the sentences.

Parameters

In information retrieval with binary classification, precision (also known as positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances retrieved. Precision and recall are therefore based on understanding and measuring relevance. In simple terms, high accuracy means that an algorithm is significantly more relevant than irrelevant results, while a high recall means that an algorithm has produced the most relevant results.

Table 1.0 Ranking Results of Reason Candidates by HTRE

CNT	Reason Candidate
353	RT @debarshi1: Promise to remove corruption was the USP of Modi's 2014 election campaign And now every single day a new scam is coming i…
220	PNB fraud ED, Nirav Modi, corruption
81	@DesiPoliticks Nothing to worry, Students can all sell Gainful Pakodas, under Modi Pakoda Rozgar Guarantee Yojna.
72	A complicit Modi Govt permits another ₹6,800 Cr bank fraud. Same modus operandi. Banks including PNB issue LOC's and fr…
31	Modi and his clan, have failed on every front and now they have failed our youth by corrupting the age-old SSC Exams. #Naukr…

IV. PERFORMANCE ANALYSIS

About Dataset:

We have been collecting live tweets for a period of 3 months on specific topics such as Narendra Modi, Donald Trump and Apple. The data has been added to the single-line value file (LSV file). To get some insights, we used the Penn Tree Bank to identify the noun: Tweets are collected with other attributes such as USERNAME, LOCATION, TWEETID, DATE-TIME, and TWEET CONTENTS. These records are then analyzed using a row separator and processed to identify variations. The implementation of heuristic grammar rules allowed improving the total gain of aspects as well as the implementation of the LDA to find subject words.

The most important category measurements for binary categories are:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$Accuracy \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$FMeaure = 2 * \frac{Precision+Recall}{Precision*Recall} \quad (5)$$

We collected 500 tweets and, using Eq. (2), (3), (4) and (5), we obtained average accuracy, recall and measurement with system accuracy. The existing system [10] has a relative accuracy of 84%, while our system offers greater than 85% accuracy.

Table 2.0 Result Analysis Key Index Parameters

Precision	Recall	Accuracy	F-Measure
85.543	99.946	85.565	92.185

CONCLUSION AND FUTURE WORK

The proposed system examines the problem of analyzing variations in public sentiment and finds the possible reasons for these variations. To solve the problem, using latent Dirichlet allocation (LDA) and Hybrid Triple Relation (HTRE) is used. Stanford's NLP is used for sentiment analysis because it provides better accuracy than existing systems. The proposed models find the variation of feeling and the mines for possible reasons behind these variations of feeling. We worked on tweets in English and rested on other languages to explore as future works. Other regional languages can be used to obtain sentiment variations, however Stanford's current PTB Tokenizer has a dictionary of three main languages other than English.

REFERENCES

[1] B. OConnor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, From tweets to polls: Linking text sentiment to public opinion time series, in Proc. 4th Int. AAAI Conf. Weblogs Social Media, Washington, DC, USA, 2010.
[2] D.M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation,"J. Mach. Learn. Res., vol. 3, pp. 993–1022, Jan. 2003.
[3] J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," in Proc.5th Int. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011.
[4] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," J. Comput. Sci., vol. 2, no. 1, pp. 1–8, Mar. 2011.

[5] Z. Wang, J. Liao, Q. Cao, H. Qi and Z. Wang, "Friendbook: A Semantic-Based Friend Recommendation System for Social Networks," in *IEEE Transactions on Mobile Computing*
[6] T. Sakaki, M. Okazaki, and Y. Matsuo, Earthquake shakes twitter users: Real-time event detection by social sensors,‖ in Proc. 19thInt. Conf. WWW, Raleigh, NC, USA, 2010.
[7] Y. Hu, A. John, F. Wang, and D. D. Seligmann, Et-lda: Joint topic modeling for aligning events and their twitter feedback,‖ in Proc.26th AAAI Conf. Artif. Intell., Vancouver, BC, Canada, 2012.
[8] D. Chakrabarti and K. Punera, Event summarization using tweets,‖ in Proc. 5th Int. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011.
[9] Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Interpreting the Public Sentiment Variations on Twitter, IEEE Transactions on Knowledge and Data Engineering, VOL. 6, NO. 1, SEPTEMBER 2012
[10] Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Interpreting the Public Sentiment Variations on Twitter,‖ IEEE Transactions on Knowledge and Data Engineering, VOL. 26, NO.5, MAY 2014.
[11] F. Liu, Y. Liu, and F. Weng, "Why is "SXSW" trending? exploring multiple text sources for twitter topic summarization," in Proc. Workshop LSM, Portland, OR, USA, 2011.
[12] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, Target dependent twitter sentiment classification,‖ in Proc. 49th HLT, Portland, OR,USA, 2011.
[13] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe,"Predicting elections with twitter: What 140 characters reveal about political sentiment," in Proc. 4th Int. AAAI Conf. Weblogs Social Media, Washington, DC, USA, 2010.
[14]B. Pang and L. Lee, Opinion mining and sentiment analysis, Found. Trends Inform. Retrieval,vol. 2, no. (12), pp 1135, 2008., vol. 14, no. 3, pp. 538-551, 1 March 2015.
[15]J. Weng and B.-S. Lee, Event detection in twitter,‖ in Proc. 5thInt. AAAI Conf. Weblogs Social Media, Barcelona, Spain, 2011.
[16] D. Hall, D. Jurafsky, and C. D. Manning, Studying the history of ideas using topic models, in Proc. Conf. EMNLP, Stroudsburg, PA, USA, 2008, pp. 363371.
[17] T. L. Gri_ths and M. Steyvers, Finding scienti_c topics, in Proc. Nat. Acad. Sci. USA, vol. 101, (Suppl. 1), pp. 52285235, Apr. 2004.
[18] G. Heinrich, Parameter estimation for text analysis, Fraunhofer IGD, Darmstadt, Germany, Univ. Leipzig, Leipzig, Germany, Tech. Rep., 2009.
[19]G. Angeli , M.Johnson, Premkumar, and C. D. Manning." Leveraging Linguistic Structure for Open Domain Information Extraction." In *Proceedings of the Association of Computational Linguistics (ACL)*, 2015.
[20] M. Hu and B. Liu, Mining and summarizing customer reviews, in Proc. 10th ACM SIGKDD, Washington, DC, USA, 2004.
[21] J. Leskovec, L. Backstrom, and J. Kleinberg, Meme-tracking and the dynamics of the news cycle, in Proc. 15th ACM SIGKDD, Paris, France, 2009.
[22] C. X. Lin, B. Zhao, Q. Mei, and J. Han, Pet: A statistical model for popular events tracking in social communities, in Proc. 16th ACM SIGKDD, Washington, DC, USA, 2010.
[23] T. Minka and J. Laerty, Expectation-propagation for the generative aspect model, in Proc. 18th Conf. UAI, San Francisco, CA, USA, 2002.
[24] G. Mishne and N. Glance, Predicting movie sales from blogger sentiment, in Proc. AAAI-CAAW, Stanford, CA, USA, 2006.
[25] A. Gupte, S. Joshi, P. Gadgul, A. Kadam, "Comparative Study of Classification Algorithms used in Sentiment Analysis," in International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014.
[26] C. D. Manning, M. Surdeanu, J.Bauer, J. Finkel, S. J. Bethard, D. McClosky, The Stanford CoreNLP Natural Language Processing Toolkit,Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, Maryland USA, June 23-24, 2014