Sentiment Knowledge Discovery in Twitter Live Streaming Data using R Language

C. Gayathri, AP/CSE
Mahendra Institute of Technology,
Mahendhirapuri, Mallasamudram,
Namakkal

R. K. Poonilavu
Mahendra Institute of Technology,
Mahendhirapuri, Mallasamudram,
Namakkal

V. Navithra
Mahendra Institute of Technology,
Mahendhirapuri, Mallasamudram,
Namakkal

A. Rajeshwari
Mahendra Institute of Technology,
Mahendhirapuri, Mallasamudram,
Namakkal

T. Asha
Mahendra Institute of Technology,
Mahendhirapuri, Mallasamudram,
Namakkal

***Abstract*—Today, social networking sites are blazing, generating and retrieving large amounts of information. Ninety-five percent of the world's population is giving their daily viewpoints on microblogging platforms, as a result of which they contain expressions that are temporary and basic. Variable gadgets, like mobile phones, laptops, tablets, and alternative IoT knowledge devices, produce large amounts of information. Net applications supported by microservices and knowledge running on these have made it easier for the US to get any variety of data at any time. It's one of the platforms for gathering massive and heterogeneous knowledge, and it's straightforward to assess human feelings through this knowledge. Sentiment analysis is the term for this. Sentiment Analysis is the task of analyzing a text's characteristics such as comments, reviews, or messages. To implement an associated formula for automatic classification of text into positive, negative, neutral, or negation. One of the most significant jobs of Natural Language Processing is sentiment analysis, often known as opinion mining. All the data employed in this graph was extracted from Twitter. To tackle this, sentiment polarity categorization is employed. Polarity of Feelings Categorization is done at the sentence and review levels.**

***Keywords*—Sentiment Analysis,Social Media,Twitter,R Programming**

I. INTRODUCTION

Today, social media is usually thought to be the most effective tool for large-scale knowledge and viewpoint sharing. The variety of information collected on an everyday basis is regularly rising because the number of users grows. This data is effective for examining people's views on current events and activities throughout the planet. Social media knowledge analysis is the method of analysing the info collected on various sites. This kind of information analysis is additionally referred to as "opinion mining," in which people's opinions and ideas are retrieved and examined, allowing a stronger understanding of people's current attitudes toward varied activities. The outstanding social networking website, Twitter, could also be utilised to fulfil the demand to know people's attitudes. Users of Twitter may send and receive "tweets", which are short communications. It's one of the platforms for gathering immense and heterogeneous knowledge, and it's straightforward to assess human feelings through the exploitation of this knowledge. Sentiment analysis is the term for this. Sentiment analysis could be a technique of extracting emotions, opinions, and reviews from varied social media platforms through the exploitation of the construct of information analysis. The document includes tweets and live comments that were retrieved using the hashtag. The report then delves into the ideas and sentiments of Twitter users, and who has commented on the live streaming. The powerful tool R was accustomed to doing the analysis of the retrieved tweets. R could be an easy and straightforward statistics programme to use and understand better.

II. RELATED WORKS

Beyond simple rating/feedback-driven recommender systems, information from online social networks has the potential to increase recommendation accuracy (RS). Many online social networks now provide "Friends Circles," a new feature that refines the domain-agnostic "Friends" concept to better suit users' activities across different domains. RS might potentially benefit from domain-specific "Trust Circles." Intuitively, a user's degree of confidence in various subgroups of friends in different domains may vary. Unfortunately, the majority of multi-category rating datasets combine all of a user's social contacts. The purpose of this research is to develop circle-based RS. We infer category-specific social trust circles using publicly available rating data and social network data. We demonstrate how to weight friends inside circles according to their assumed competency levels in a number of ways Through testing on publically available data, we show that the recommended circle-based recommendation models may better use users' social trust information, resulting in increased recommendation accuracy.Twitter is a valuable instrument for tracking and gauging public opinion since it contains millions of users who voice their opinions. Such surveillance and analysis can provide vital information for decision-making in a range of domains. As a result, both academics and industry are interested in it. Previous research has mostly focused on public mood tracking and modelling.

In this study, we take it a step further by evaluating sentiment variations. We observed that emerging themes (called foreground subjects) are substantially linked to the true reasons for the differences during the sentiment variation periods. We propose the Foreground and Background LDA (FB-LDA), a Latent Dirichlet Allocation (LDA)-based model for distilling foreground concerns and filtering out long-standing background themes, based on this discovery. These primary topics can assist in interpreting mood fluctuations To improve the readability of the mined reasons, we use a generative model called Reason Candidate and Background LDA to pick the most representative tweets for foreground subjects (RCBLDA).

According on experimental data, our systems can successfully recognise foreground themes and rate reason candidates. The models given here might be used, for example, to discover subject differences between two sets of texts.

The increase of online opinion data has forced the development of automated tools to analyse and interpret people's sentiments on a variety of topics. In most sentiment analysis systems, the sentiment lexicon is key. There is no generally optimal emotion vocabulary, however, because the polarity of words is sensitive to the subject area. Worse, the same phrase might be used to describe several polarities within the same issue. For example, in a laptop review, the adjective "large" is negative for the battery but good for the screen. We investigate the difficulty of learning a sentiment lexicon from an unlabeled opinionated text collection that is not only domain specific but also dependent on the aspect in context in this study. We provide an original optimization technique for building a context-dependent sentiment lexicon that provides a rational and consistent framework for incorporating numerous sources of data Experiments on two data sets (hotel reviews and customer feedback surveys on printers) show that our system can not only identify new emotion words unique to a domain, but also distinguish between many polarities of a term based on its context. We were able to show that our method is successful in developing a high-quality vocabulary by comparing it to a human-annotated gold standard. Furthermore, using the learned context-dependent sentiment lexicon improved the aspect-level accuracy of a sentiment classification job.

III. EXISTING SYSTEM

The Support Vector Machine is used for sentiment analysis based on categorization because it delivers the highest level of sentiment accuracy. It's a classification method for both linear and nonlinear knowledge. The SVM looks for the best separating hyperplane that is linear (the linear kernel). That might be a call border separating one category's knowledge from another. The SVM employs nonlinear mapping to rework the information into a better dimension if the input is linearly indivisible. It then finds a linear hyperplane to solve the problem. Biased reviews, subtlety, thwarted expectations, ordering effects, characteristics or traits, and difficult interpretation of the resulting model are all drawbacks of the current approach.

IV. PROPOSED SYSTEM

The sentence-level categorizer is used to aggregate the datasets from Twitter in this case. The tokenizer then tokenizes the datasets. After that, knowledge processes the tokens. Knowledge cleansing, knowledge integrity, and knowledge transformation are all linked to knowledge. Reduction is assigned to a distinct format. Whether the provided data is positive, negative, neutral, or negation, this might be implied in a sign for characteristic. The datasets are classified using the Naive Bayes Classifier, which is the most successful classifier for sentence level classification. The R-platform visualises the classified data using the Naive Mathematical Classifier formula. The projected system has the subsequent advantages. A model is straightforward to interpret, domain-specific, and involve a lot of strong, economical computation.

The design diagram given below explains the method:

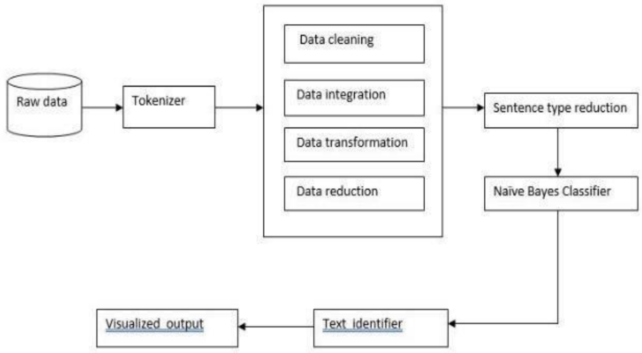

Fig.1. Design Diagram

V. WORKFLOW

A. *Fetching Data*

We must obtain raw data from Twitter in order to do analysis using the R programming language in this project. By connecting to Twitter's Stream API, the Stream R package allows users to get real-time Twitter data.

B. *Tokenizing*

In emotional analysis, tokenization is the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements known as tokens. Following that, the token list is utilised as input for parsing or text mining. Tokenization has uses in both linguistics and computer science (as a type of text segmentation) (as part of sentiment analysis). A tokenizer takes a stream of characters and breaks them down into individual tokens (usually words) before generating a stream of tokens.

C. Data Pre-processing

Data pre-processing is a data mining technique that involves converting raw data into a usable format. Real-world data is frequently insufficient, inconsistent, or missing in specific behaviours or patterns, and it contains numerous inaccuracies. Pre-processing data is a tried-and- true method of dealing with such problems. Data preprocessing is the process of preparing raw data for further processing.

D. Sentence Detection

Following the completion of the data, The data is translated into a comprehensible format during pre-processing. This ensures that the data is correctly identified and understood. In addition, the sentence is reduced to a declarative, imperative, and interrogative statement..

E. Classifying the text

The Naive Bayes classifier is used to categorise the data in this project. The Naive Bayes classifier is a machine learning classifier that categorises data in certain ways, such as positive or negative data. Naive bayes is a basic classification technique that relies on strong assumptions about each input variable's independence.

F. Visualizing the Data

It generates data classification outputs such as positive, negative, neutral, and negation, which are calculated from Twitter.

VI. PERFORMANCE

We haven't used the Twitter API to analyse tweets received over the site yet. That will be done via live broadcasting. In addition, we'll use sentiment analysis to determine if individuals are favourable, negative, or neutral about the material.

A. Log in to the Twitter API after completing the authentication process.

If you want to use the Twitter API, you'll need to develop an app and authorise it first. Keep in mind that you should only share your keys with people you trust with your account.

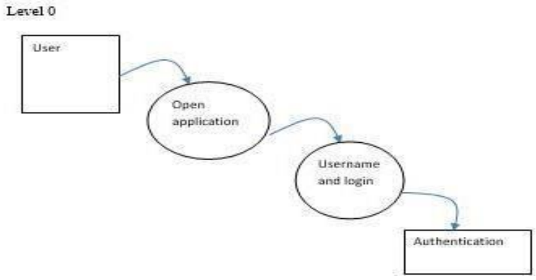

Fig.2. Data Flow Diagram at Level 0

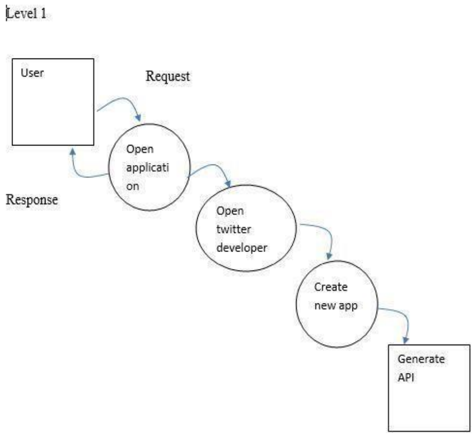

Fig.3. Data Flow Diagram at Level 1

B. Gather some tweets

Collect the necessary number of tweets for analysis.

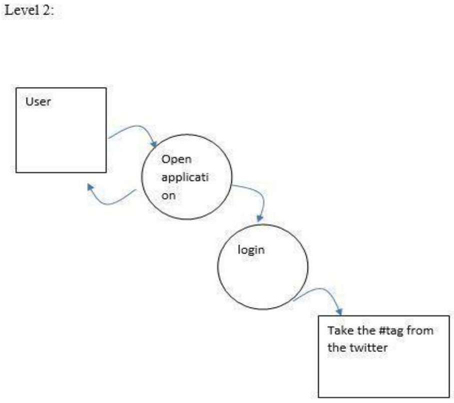

Fig.4. Data Flow Diagram at Level 2

C. Sentiment analysis

The science of assessing whether words are favourable or negative is known as sentiment analysis. The goal is to determine the overall tone of a document by examining the meanings of the words that make it up. Then, using the Naive Bayes Classifier Algorithm, sort them into positive, negative, or neutral categories. By comparing each individual word to the lexicon, we may get a general idea of whether the assertions are good or negative. These favourable words receive a four or five out of five rating. Negative words such as "hate," "poison," and "awful" receive negative scores of minus five. There are several classifications you may use instead of this one to gauge emotion. The first step is to break down the tweets into words.

D. Merge the Twitter data with the sentiment scores

This creates a data frame with each tweet's term as a new row. We'll now combine it with the lexicon. Any terms that don't match the database will be removed, including hashtags, @ handles, and other missing words. We're almost ready to graph after a little more polishing.

E. Graph and Visualize Charts

We adjusted each tweet's timestamp to the closest hour in the previous step. A sentiment value, a timestamp, and the nearest hour are all included in each entry. Because the word and the person who first tweeted it are no longer relevant, we no longer need them. The next step before graphing is to add together each hour's mean net sentiment score. Initially, I averaged the sentiment values for each hour, but I discovered that it correlated strongly with the number of tweets, making it useless.

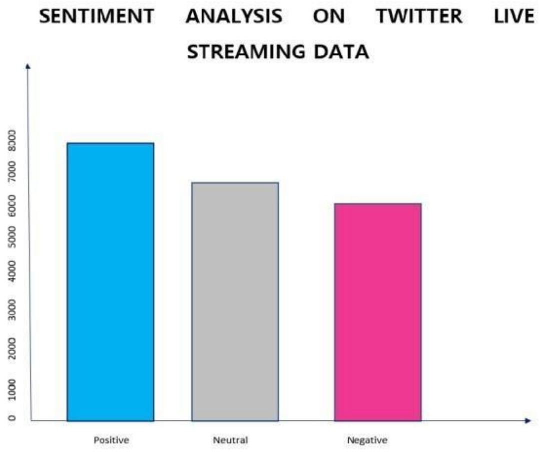

Fig.5. Result Model

VII. CONCLUSION

Sentiment analysis is a nebulous but valuable science. Sentiment analysis, often known as opinion mining, is a popular topic these days. Because of the complexity of the English language, and much more so when considering other languages such as Chinese, we are still a long way from reliably detecting the moods of a corpus of writings. We attempted to demonstrate the fundamental method of categorising tweets into positive or negative categories using Naive Bayes as a baseline, as well as how language models are connected to Naive Bayes and may give superior results. We might enhance our classifier even further by extracting additional features from the tweets, experimenting with various types of features, tweaking the Naive Bayes classifier's parameters, or attempting a new classifier altogether. I'm sure you've realised the importance of sentiment analysis by now. Sentiment analysis might be used to a far broader range of situations, including photographs. Even though there are several tools on the market, having a practical understanding of how the entire process works is advantageous. Furthermore, the current tools are prohibitively costly and do not provide the kind of flexibility and customization that R allows.

VIII. REFRENCES

[1] X. Yang, H. Steck, and Y. Liu, "Circle-based recommendation in online social networks, " in Proc. 18th ACM SIGKDD Int. Conf. KDD, New York, NY, USA, Aug. 2018, pp. 1267–1275.

[2] S. Tan, Y. Li, H. Sun, Z. Guan, X. Yan, "Interpreting the public sentiment variations on twitter," IEEE transactions on knowledge and data engineering, vol. 26, no. 5, may 2019, pp. 1158- 1170.

[3] G. Zhao, X. Qian, Xie, "User-service rating prediction by exploring social users' rating behaviors," IEEE Transactions on Multimedia, 2019, 18(3):496-506.

[4] Y. Lu, M. Castellanos, U. Dayal, C. Zhai, "Automatic construction of a context-aware sentiment lexicon: an optimization approach," World Wide Web Conference Series. 2019, pp. 347-356

[5] X. Qian, H. Feng, G. Zhao, and T. Mei, "Personalized recommendation combining user interest and social circle," IEEE Trans. Knowledge and data engineering. 2018, pp. 1763-1777.

[6] M. Taimoor Khan, Shehzad Khalid, "Sentiment Analysis for Health Care", International Journal of Privacy and Health Information Management, 2015.

[7] Eman M.G. Younis, "Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study", International Journal of Computer Applications, Volume 112 – No. 5, February 2015.

[8] Akshi Kumar and Teeja Mary Sebastian, "Sentiment Analysis on Twitter", International Journal of Computer Science Issues, Volume 9, Issue 4, No. 3, July 2012.

[9] Pierre Ficamos, Yan Liu, "A Topic based Approach for Sentiment Analysis on Twitter Data", International Journal of Advanced Computer Science and Applications, Vol. 7, No. 12, 2016.

[10] Pooja Khanna, Sachin Kumar, Sumita Mishra, Anant Sinha, "Sentiment analysis: An approach to opinion mining from twitter data using R", International Journal of Advanced Research in Computer Science, Volume 8, No. 8, 2017.

[11] Kiruthika M., Sanjana Woonna, Priyanka Giri, "Sentiment Analysis of Twitter Data", International Journal of Innovations in Engineering and Technology, Volume 6, Issue 4, April 2016.

[12] Shubham S. Deshmukh, Harshal Joshi, Pranali Pandhare, Aniket More, Prof.Aniket M. Junghare, "Twitter DataAnalysis using R", International Journal of Science, Engineering and Technology Research, Volume 6, Issue 4, April 2017.

[13] Varsha Sahayak, Vijaya Shete, Apashabi Pathan, "Sentiment Analysis on Twitter Data", International Journal of Innovative Research in Advanced Engineering, Issue 1, Volume 2, January 2015.

[14] Sonal Singh, Shyam S Choudhary, "Social Media Analysis: Sentiment Analysis Twitter Using R Language", International Journal of Advances in Electronics and Computer Science, Volume 4, Issue 11, November 2017.

[15] Onam Bharti, Mrs. Monika Malhotra, "Sentiment Analysis", International Journal of Computer Science and Mobile Computing, Volume 5, Issue. 6, pages 625 – 633, June 2016.

[16] Sudipta Roy, Sourish Dhar, Arnab Paul, Saprativa Bhattacharjee, Anirban Das, Deepjyoti Choudhury," Current Trends of Opinion Mining and Sentiment Analysis In Social Networks", International Journal of Research in Engineering and Technology, Volume 2, Special Issue 2, December 2013.

[17] Federico Neri, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, Tomas By, "Sentiment Analysis on Social Media ", Published at ACM International Conference on Advances in Social Networks Analysis and Mining, 201