OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

REVIEW PAPER ON SENTIMENT ANALYSIS STRATEGIES FOR SOCIAL MEDIA

SHAIKH SAYEMA ANWER[1] , V. S. KARWANDE[2]

Department of Computer Science & Engineering, EESGOI , India[1]

HOD, Assistant Professor, Department of Computer Science and Engineering, EESGOI, India. [2]

Abstract: ***Natural language processing is used in sentiment analysis. It's also known as sentiment analysis or opinion mining. It aids decision-making in humans. Various tasks, such as subjectivity identification, sentiment classification, aspect term extraction, feature extraction, and so on, are required to perform sentiment analysis. Users can easily express their thoughts and feelings by using social media sites such as Twitter, Facebook, and others. Millions of people share their views through their everyday interactions on social media sites such as Twitter, Facebook, and others, which can be their sentiments and opinions about a specific subject. These ever-increasing subjective data are unquestionably a wealth of knowledge for any type of decision-making process. Sentiment Analysis is a field that has arisen to automate the analysis of such results. Its aim is to recognize data on the Internet and classify it according to its polarity, or whether it has a positive or negative connotation. Sentiment Analysis is a text-based analysis issue, but it has some problems that make it more challenging than conventional text-based analysis.***

Keywords: *sentiment analysis, sentiment classification, features selection, opinion*

I INTRODUCTION

Sentiment analysis, also known as opinion mining, is a method for automatically finding opinions embodied in text that is becoming a challenge in many research fields, especially in the data mining field for social media, with a variety of applications such as product reviews, input analysis, and consumer decision making, among others. The method of extracting feelings or thoughts from a piece of text for a specific subject is known as sentiment analysis. It enables us to comprehend the text's attitudes, thoughts, and feelings. It gathers information about a user's likes and dislikes from web content. It entails predicting or analyzing the text's concealed content. This secret data is extremely useful for gaining insight into a user's preferences and dislikes. The aim of sentiment analysis is to figure out what a writer or speaker thinks about a particular subject. Audio, photographs, and videos can all be used for sentiment analysis. The internet has now become an integral part of our daily lives. The majority of people use online blogging or social networking sites to share their thoughts on various topics. They often use these forums to learn about other people's viewpoints. As a result, data mining and sentiment extraction have become relevant research areas. As HUMANS, we are drawn to those who share our values. Even studies show that we feel more at ease socializing with people who share our values, with people we can trust and who can assist us in achieving our goals. People have an etymological propensity to identify with like-minded groups. A society is made up of several clusters. Modularity is one of the most important factors to consider when deciding the number of cultures. If the clusters' characteristics are thoroughly examined, it may aid in identifying the unique character set of individual clusters or groups of like-minded individuals. To put it another way, the existence of a shared connect with a group of people means that those individuals share similar values and goals. To be more precise, two forms of social media are available: Social Networks and Online Communities. People who are linked by previous personal connections form social networks, which they maintain socially and would like to link to new associations to expand their personal contacts. It connects people who have a direct line of communication with one another. In contrast to the former, cultures are made up of individuals from various fields who have little or no common ground. The fondness for a familiar interest serves as the primary link between

members of a society. People tend to remain within a group for a variety of reasons, including a fondness for a particular item, a sense that he or she should be associated with that community, or the possibility of achieving something by adhering to that community. Social networks clearly contain organized arrangements, while groups contain overlapped and nested arrangements. Social media is a means of disseminating information to a large and diverse audience. It can be thought of as a means of disseminating knowledge through a GUI. Individuals can tailor their content to a broader society and reach out to more people for sharing or promotion using social media in conjunction with social networks. The process of categorizing the opinions expressed about a specific object is known as sentiment analysis. With the emergence of various technological tools, it has become critical to be mindful of the general public's opinion in matters of industry, goods, and common likes and dislikes. Identifying the emotion behind social media posts will assist in determining the context in which the individual can respond and improve.

II LITERATURE SURVEY FOR SENTIMENT ANALYSIS

2.1. SENTIMENT CLASSIFICATION

Sentiment analysis, also known as "opinion mining" or "emotion Artificial Intelligence," refers to the systematic recognition, extraction, evaluation, and examination of emotional states and subjective information using natural language processing (NLP), text mining, computational linguistics, and bio measurements. Sentiment analysis is mostly concerned with the voice of client content, such as polls and feedback on the Internet and social media sites.

- Document-level sentiment classification: A document may be completely categorised as "positive," "negative," or "neutral" at this level.
- Sentence-level emotion classification: Each sentence is categorized as "positive," "negative," or "neutral" at this level. Sentences/documents may be classified as "positive," "negative," or "non-partisan" at this level, based on some aspects of the sentences /archives, and this is widely referred to as "perspective-level appraisal grouping."

DIFFERET ALGORITHMS ARE USE IN SENTIMENT ANALYSIS :

2.1.1 A classifier based on the Naive Bayes algorithm.

It's a Bayes' theorem-based algorithm that uses a small amount of training data to estimate the parameters, also known as probabilistic classifiers. It is widely regarded as the fastest classifier, as well as the most scalable, and it can handle both discrete and continuous data. This algorithm was used to make a real-time prediction. There are various types of naive classifiers, including multinomial classifiers. Nave Bayes, Bernoulli Nave Bayes, and Gaussian naive are all terms for the same thing. formula for classification algorithms :

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

The following is a Bayesian description of posterior probabilities:

P(A|B)- Posterior probabilities, where A and B are cases.

If two values are unrelated to one another, then P(A, B) = P(A) (B)

The python library can be used to create Naive Bayes. Through they are used in recommendation systems, Naive's predictors are independent. They are well-known in document classification and are used in many real-time applications.

2.1.2 2.2.2. Decision tree

It's a top-down model with a flow-chart structure that handles high-dimensional data. Based on the given input variable, the outcomes are predicted. The following elements make up a decision tree: There's a root, a lot of nodes, branches, and leaves. The root node partitions the nodes based on the class's attribute value, the internal node takes an attribute for further classification, the branches use a decision rule to split the nodes into leaf nodes, and the leaf nodes provide the final result. The decision tree's time complexity is determined by the number of records and training data attributes. It's difficult to get the desired results if the decision tree is too long.

Advantage: They are used in day-to-day activities to choose the target based on decision analysis and for predictive analytics to solve problems. Creates a model based on the source data automatically. Best at dealing with missing values.

Disadvantage: The tree's size is uncontrollable until it reaches some sort of limit. Trees are unstable due to their hierarchical structure.

2.2.3. Support Vector Machine.

This algorithm is a machine learning supervised algorithm that is used to solve classification problems. It's a crucial tool for researchers and data scientists. The method of this SVM is to find a hyperplane in an N-dimensional space of data points. Decision boundaries that distinguish data points are known as hyperplanes. This vector falls closer to the hyperplane, maximizing the classifier's margin. The generalization error is lowest when the margin is full. Their implementation can be achieved using python and some

training datasets with the kernel. The SVM's main goal is to train an entity to fit into a specific classification. The SVM isn't limited to being a linear classifier. Because of their kernel function, which increases computational performance, SVM is favors over any other classification model.

Advantage: They are preferred due to their low computing power and high accuracy. Good memory performance and effective in high-dimensional space.

Limitations in rpm, kernel, and size are all disadvantages.

2.2.4. Random Forest.

It's a sophisticated machine-learning algorithm focused on Ensemble learning. The decision tree, which is used to construct predictive models, is the most basic building block of Random forest. The pruning method is done by setting a stopping splits to produce a better outcome, and the job demonstration involves constructing a forest of random decision trees. For decision-making, Random Forest employs a methodology known as bagging. By reducing the bias, this bagging avoids overfitting of data, and this random can achieve better accuracy. An average of several decision trees, or frequent forecasts, is used to make a final prediction. The random forest has a wide range of applications, including stock market forecasting, fraud detection, and news forecasting.

Advantages include: It doesn't take a lot of computing power to process the datasets, and it's a simple model to construct. Greater precision aids in the resolution of predictive issues.

It handles missing values well and detects outliers automatically.

Negative aspects: Requires a lot of memory and a lot of computation.Requires a significant amount of time.

2.2.5 K- Nearest Neighbors

For CART, we'll look at the K-NN algorithm with supervised learning. They use a K positive small integer to assign an individual to a class based on its neighbours, or assigning a group by observing the group the neighbour belongs to. This is determined using the Euclidean distance and brute force. The Tuning procedure can be used to determine K's value. KNN prefers to use normalisation to rescale data rather than learning any model to train a new dataset.

Advantage: If the training data is large, it produces successful results.

The main disadvantage is that it only works well if the variable is small. Second, when classifying, choose the K factor.

III SYSTEMS ARCHITECTURE

Text summarization is a type of data mining that uses natural language processing to extract people's opinions from a variety of businesses and shopping sites. The recent internet trend of encouraging users to contribute their opinions and suggestions has resulted in a massive collection of useful information. The opinion mining system examines each text to determine which sections contain opinionated words, which sections are being opinionated, and who wrote the opinion. Sentiment analysis decides the sentiment polarity orientation of each opinionated word or expression, whether it is positive, negative, or neutral. It expresses a writer's or speaker's opinion in a concise manner. At the sentence stage, text summarization is performed. To derive the exact opinion of customers about the commodity, the proposed framework employs a text summarization technique.

The overall design of the proposed structure is depicted in the diagram below. Taking user input, POS tagging, sentence compression, using a naive bayes classifier algorithm, and polarity detection are all phases of the architecture. Each phase serves a distinct purpose. These functionalities described as following:

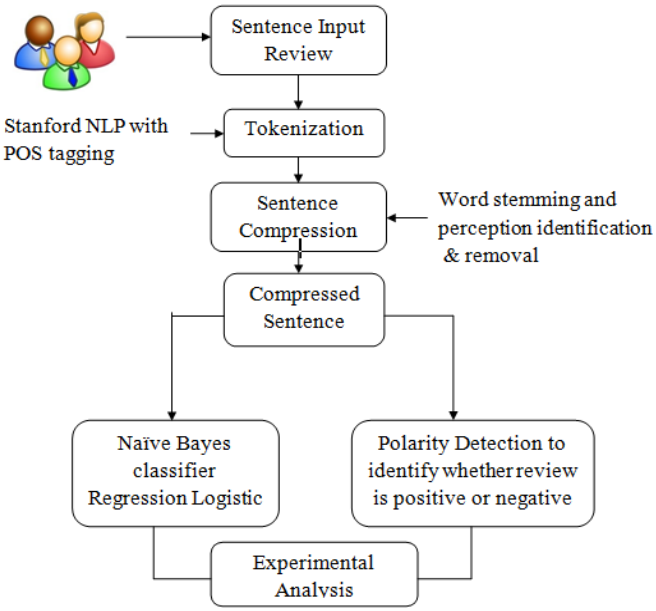

Figure No.3.1: Proposed System Architecture

A. User Input: It accepts user input: The term "input" refers to user reviews or comments on a specific product. Users have the option of posting reviews with or without product selection. Because the system is supervised to classify product reviews into their appropriate categories, whether they are given by selecting the product or not.

B. Tokenization is the second step:

Tokenize - Turns a string of text into a series of tokens. The English component includes a PTB-style tokenizer that has been extended to handle noisy and web text reasonably well. A tokenizer breaks down text into tokens, which are roughly equivalent to "words." PTBTokenizer is a class that is suitable for tokenization of English. It was created to closely resemble Penn Treebank (PTB) tokenization, hence the name and the fact that it is provided by the Stanford Natural Language Processing Group (Standford NLP). This tokenization is also used by an auxiliary tool to split text into sentences. PTBTokenizer is more focused on formal English writing than SMS-speak. Standford NLP tokenixer was used to tag parts of speech (POS). Tokens are classified into grammatical features such as noun, adjective, and adverb using POS tagging.

C. Stemming, Perception words and cardinal digits removal :

In order to reduce each word to its root form, stemming attempts to eliminate the distinctions between inflected forms of a word. For example, foxes could be reduced to the root fox to eliminate the distinction between singular and plural in the same way that lowercase and uppercase were eliminated. The canonical, or dictionary, form of a group of similar terms is a lemma; for example, the lemma of paying, paid, and pays is pay. The lemma of is, was, am, and being is be; nevertheless, the lemma of is, was, am, and being does not always imitate the terms it is compared to. Lemmatization, like stemming, attempts to group similar words together, but it goes a step further by attempting to group words according to their word sense, or context. A single word may have several meanings, such as wake, which can refer to both waking up and a funeral. Although lemmatization tries to differentiate between these two word senses, stemming conflates them incorrectly. Lemmatization is a much more difficult and costly process that requires an understanding of the context in which words appear before making decisions on what they mean. Stemming tends to be just as successful as lemmatization in nature, but at a lower cost.

It's possible that a word's root form isn't even a real word. Both the terms jumping and jumpiness can be derived from the word jumpy. It makes no difference as long as the same terms are produced at index and search time; search will simply work. There would be only one implementation if stemming was easy. Unfortunately, stemming is an imprecise science with two flaws: under stemming and over stemming. Failure to reduce terms of the same meaning to the same root is known as under stemming. Jumped and jumps, for example, could be reduced to jump, while jumping could be reduced to jumpi. Under stemming limits retrieval by omitting related records. The inability to keep two terms with distinct meanings apart is known as over stemming. For example, both general and generate could be stemmed to gener. Over stemming decreases accuracy by returning irrelevant records when they shouldn't be. I believe, I sound, luckily, and other perception terms used by users in comments or reviews.

As a result, those perception terms that do not alter the context or emotion orientation of the original sentence must be deleted. Because of this sentence compression method, only the simple root words and those words that accurately represent the users' opinions are kept. When writing a review or article, users often use cardinal digits such as +4,-3, and so on. As a result, such digits pose a problem when it comes to categorizing terms. The proposed scheme eliminates certain cardinal digits, leaving only the terms to be classified.

IV CONCLUSION

The proposed framework employs a set of data mining and natural language processing techniques to summarise product feedback. The process of generating a grammatical description of a single sentence with minimal information loss is referred to as automatic sentence compression. It has recently gained a lot of interest, partly due to its applicability. The proposed structure analyses product feedback to determine their polarities (positive, negative, and neutral). The system makes use of a database of electronic objects such as computers, cellphones, cameras, and so on. Only the English language is supported by the proposed scheme. The Navie Bayes classifier essentially divides compressed sentences' polarity terms into two categories: positive, negative, and neutral.

REFERENCES

1. Che, Wanxiang, Yanyan Zhao, Honglei Guo, Zhong Su, and Ting Liu. "Sentence Compression for Aspect-Based Sentiment Analysis." Audio, Speech, and Language Processing, IEEE/ACM Transactions on 23, no. 12 (2015): 2111-2124
2. David Zajic1, Bonnie J. Dorr1, Jimmy Lin1, Richard Schwartz. “Multi-Candidate Reduction: Sentence Compression as a Tool for Document Summarization Tasks.” University of Maryland College Park, Maryland, USA , 2BBN Technologies 9861 Broken Land Parkway Columbia, MD 21046.
3. Kim, Soo-Min, and Eduard Hovy. "Automatic identification of pro and con reasons in online reviews." In Proceedings of the COLING/ACL on Main conference poster sessions, pp. 483-490. Association for Computational Linguistics, 2006.
4. Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." In Proceedings of the tenth ACM

SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177 ACM, 2004.

5. Jin, Jian, Ping Ji, and Ying Liu. "Translating online customer opinions into engineering characteristics in QFD: A probabilistic language analysis approach." Engineering Applications of Artificial Intelligence 41 (2015): 115-127.
6. Suanmali, Ladda, Naomie Salim, and Mohammed Salem Binwahlan. "Fuzzy logic based method for improving text summarization." arXiv preprint arXiv:0906.4690 (2009).
7. Ghorashi, Seyed Hamid, Roliana Ibrahim, Shirin Noekhah, and Niloufar Salehi Dastjerdi. "A frequent pattern mining algorithm for feature extraction of customer reviews." In IJCSI International Journal of Computer Science Issues. 2012.
8. Mrs. Elakkiya.R, Mrs. Jayasudha.M2, Mr. Sivanesh Waran."Improved Optimized Sentiment Classification On Dynamic Tweets".IJCSMC, Vol. 5, Issue. 6, June 2016, pg.11 – 22,ISSN 2320–088X IMPACT FACTOR: 5.258.
9. LuWang,Hema Raghavan,Vittorio Castelli ."A Sentence Compression Based Framework to Query-Focused Multi-Document Summarization".Cornell University, Ithaca, NY 14853, USA ,T. J. Watson Research Center, Yorktown Heights, NY 10598, USA.
10. Alejandro Molina1, Juan-Manuel Torres-Moreno, Eric SanJuan,Iria da Cunh, and Gerardo Eugenio Sierra Martínez,LIA."Discursive Sentence Compression".Université d'Avignon,IULA-Universitat Pompeu Fabra,GIL-Instituto de Ingeniería UNAM.
11. Kapil Thadani and Kathleen McKeown. "Sentence Compression with Joint Structural Inference".Department of Computer Science Columbia University New York, NY 10025, USA.
12. Dipanjan Das Andre, F.T. Martins. "A Survey on Automatic Text Summarization" Language Technologies Institute Carnegie Mellon University ,November 21, 2007.
13. Seyed Hamid Ghorashi, Roliana Ibrahim, Shirin Noekhah and Niloufar Salehi Dastjerdi. "A Frequent Pattern Mining Algorithm for Feature Extraction of Customer Reviews" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 1, July 2012 ISSN (Online): 1694-0814.
14. Kevin Knight,Daniel Marcu. "Summarization beyond sentence extraction: A probabilistic approach to sentence compression". Information Sciences Institute and Department of Computer Science, University of Southern California,4676 Admiralty Way, Suite 1001, Marina del Rey, CA 90292, USA Received 11 May 2001.
15. Trevor Cohn and Mirella Lapata. "Sentence Compression Beyond Word Deletion". Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 137–144 Manchester, August 2008.