International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Semantic Analysis on Social Media

Seerat choudhary[1], Jyoti Godara[2]

[1]*Student, Department of Computer Science and Engineering, Lovely Professional University, India*

[2] *Assistant Professor, Department of Computer Science and Engineering, Lovely Professional University, India*

Abstract — Social media sentiment analysis allows to track the opinion about the current ongoing topics and other issues quickly and easily. Several features and techniques for training sentiment classifiers on datasets have been studied in recent years, with inconsistent results. In this paper, we offer a technique for detecting emotion in text and predicting sentiment utilizing semantics as additional features, such as hashtags, which are widely used nowadays in texts. The intelligent technology called as naive bayes is presented in this research for Opinion Mining on social media. It introduces a multinominal naive bayes strategy for extracting semantic and affective components of semi-structured and unstructured text resources from social media sites such as Twitter. This strategy will make the analysis more efficient. This approach will improve the efficiency of the analysis on social media content or the specific data and fuzziness of natural language. The assorted techniques like bag of words, TF-IDF for converting data into features is proposed.

***Index Terms*— sematic analysis, hashtags, multinominal naive bayes, social media.**

I. INTRODUCTION

Semantic analysis is opinion expressed in a piece of text or emotions to determine whether the topic, product, movie, etc is positive, negative or neutral. These helps data analysts within larger enterprise about the public opinion, monitor brands and product reputation in market. There are many technique to determine semantic analysis it can be through Fine-Grained sentiment, Emotion Detection, Intent analysis, Aspect-based analysis, etc. A Sentiment Analysis tool is used to analyze text conversation and evaluate the tone, intent and emotion of the message. It determines the text in positive, negative or neutral and to analyze these Natural Language processing and Machine Learning technique is combined to assign score to the text, application and categories. With the growth of social networks, an increasing number of people want to find, and exchange information with each other without any regard to the geographical distance. Seen As a result, individuals want rapid, free, and easily accessible technologies to assist them in meeting these demands. Users' requirements can be met through social networks. Every day, the number of people using social media grows, and they tend to share what they know about the issues that interest them. If we know how to use Twitter's data effectively, it may give a wealth of benefits.

Fig 1. Emotional Rating

1.1 Social Media Importance –

Global interacting may have taken the world by astonishment, reducing the world's aspects. There is a good amount of contact between people via various social media networks such as Twitter. In

furthermost cases, these dialogues would be informal, and they would reflect the tone and attitude of those engaging in the argument. This opens the door to considering the participants' behavioral tendencies in the dispute. Sentimental research, also known as techniques, may be employed in the analysis and interpretation of transcript interchange to identify and discard obsolete and unneeded data features that do not add to the accuracy of a prediction model, potentially lowering the model's accuracy. There are less features that are beneficial since the model's difficulty is reduced, and it is calmer to recognize and define a simpler model.

Fig 2. Networking on social media

1.2 Machine Learning -

Machine learning is a sort of artificial intelligence that enables software programs to anticipate results without being explicitly designed to do so. Machine learning algorithms simulate new output values using past data as feedback. A reference engine is a common machine learning application. Sensing scam, junk cleaning, malware spasm recognition, corporate procedure mechanization, and logical protection are some of the other common applications.

1.3 Fine-grained sentiment -

This study can help you comprehend the input you receive from clients. In terms of the polarity of the input, you can achieve exact results. However, as compared to other sorts of research, the process of understanding this might be more time-consuming and costly.

1.4 Emotion Detection Sentiment Analysis -

This is a more advanced method of detecting emotion in a piece of writing. The sentiment can be shown using lexicons and machine learning. Lexicons are collections of positive and negative words. This makes it easy to categories the phrases based on their emotional content. The benefit of doing so is that it allows a firm to understand why a consumer feels a certain way. This is more algorithm-based, and therefore may be difficult to grasp at first.

1.5 Aspect-based –

This method of sentiment analysis is usually used to assess one component of a facility. If a organization that sells microwave services sentiment analysis, it might be for one element of televisions, such as brightness, sound, and so on. So that they can learn how customers feel about certain product features.

1.6 Naive bayes classifier –

The Naive Bayes technique is a directed learning technique for addressing organization issues that is constructed on the Bayes formula. It is mostly utilized in text organization tasks that need a large training datasets. The Naive Bayes Classifier is a modest and effective organization technique that aids in the development of rapid machine learning models capable of making quick predictions. It's a probabilistic classifier that is, it makes predictions based on an entity's likelihood. Junk cleaning, sentiment analysis, and article classification are all common uses of the Naive Bayes Procedures.

Types of naive bayes model:

There are basically three kinds of Naive Bayes Model, such as mentioned below:

Gaussian: It implies that characteristics are distributed normally. If analysts accept constant values rather than distinct values, the model assumes that these values are drawn from a Gaussian dissemination.

Multinomial: When the data is multinomial distributed, it is utilized and it is generally used to solve text categorization issues, which involves determining which group a text goes to, such as Sports, Politics, or Learning. The predictions in the classifier are based on the occurrence of terms.

Bernoulli: The classifier operates in a related way to the Multinomial classifier, excluding that the interpreter variables are autonomous Booleans variables. For instance, defining whether or not a specific term seems in a text. This example is also recognized for jobs concerning text classification. The thousands of programs or tools for analyzing arithmetic data, but only a handful for analyzing words.

The Multinomial Naive Bayes algorithm is a probabilistic learning approach popular in Natural Language Processing. The package guesses the tag of a manuscript, like a message or a broadsheet story, using the Bayes formula. It evaluates the likelihood of each label for a likely example and then outputs the label with the greatest possibility. A Naive Bayes classifier is gathering of various procedures that follows to a mutual standard: each article being classed is unrelated to any further article. The article's existence or absence has no bearing on the presence or lack of another article.

II. LITERATURE REVIEW

The author of the work [1] tackled the issue of applying the classic naive Bayes model to the categorization of uncertain data. A session provisional prospect estimate is a critical problematic in the naive Bayes model, and kernel density estimation is a popular technique for doing so. The kernel density estimate technique has been improved to cope with unclear data. The problem is reduced to the examination of double-integrals as a result of this. Extensive studies on many UCI datasets indicate that using the complete pdf information of uncertain data in the uncertain naive Bayes model may yield classifiers with greater accuracy than expending the mean as the characteristic value of indeterminate facts. Interval difficulty study and experiment-based presentation study show that the formula-based technique has distinct benefits over other approaches.

The naïve Bayes classifier makes learning even easier by assuming that characteristics are consistent within a given class. However, while independence is typically a bad assumption, naive Bayes may occasionally match up well with more advanced classifiers in practice. [2] Despite its implausible independence assumption, the nave Bayes classifier is surprisingly successful in practice, since its classification choice is typically right even if its probability estimations are incorrect. Although several naive Bayes optimality requirements have been found in the past, a better knowledge of the data properties that influence naive Bayes performance is still needed. Instead, when using a naïve Bayes model, the information loss that features include the class is a stronger predictor of accuracy.

An auxiliary feature approach is presented in the study [3]. It chooses an auxiliary feature that can reclassify the text space directed towards the chosen features using an existing feature selection approach. The relevant conditional probability is then tweaked to improve classification precision. The suggested approach does indeed increase the performance of the naive Bayes classifier, as demonstrated by examples.

Citations acknowledge the influence of scientific publication has on subsequent work. And also,

deciding how and when to cite a paper, is heavily influenced by social factors.

In paper [4], they have conduct an empirical analysis grounded on a dataset of 2010–2012 global publications in chemical engineering. They used a social network analysis and text mining to measure publication attributes and understand which variables can better help predicting their future success. Controlling for intrinsic quality of a publication and for the number of authors in the byline, it is predicted that scholarly impact of a paper in terms of citations that was received in 6 years after publication is nearly with 80% accuracy. Resulted in better to co-publishing with rotating co-authors and write the papers abstract using more positive words, and a more complex, thus more revealing, language. Publications resulted in the collaboration of different social groups also appeal more citations.

This paper [5] presents a novel statistical technique for factor analysis of binary and count data which is associated to a technique known as Latent Semantic Analysis.

The semantic image segmentation comprises of classifying each pixel of image into an instance, where each instance corresponds to the class. This idea of scene better explains the global context of an image. In medical image analysis area, image segmentation is applied for image-guided interventions, radiotherapy, or improved radiological diagnostics. In this review[6], they categorize the leading deep learning-based medical and non-medical image segmentation solutions into six main groups of deep architectural, data synthesis-based, loss function-based, sequenced models, weakly supervised, and multi-task techniques and provides complete review of the contributions in each of these groups. Further each group analyzes each variation of these groups and discuss the limitations of the current approaches and present potential future research directions for semantic image segmentation.

Sentiment calculation helps some new request chances and method tests in artificial intelligence for next group, and it has established a fascinating research field. In this paper [7] Wang, Y., Rao, Y. proposed the conception of sentiment computing with some core elements, feature vectors and some vital issues. This subjective content or objective content is classified by some special algorithms in the scenarios of single modal, such as text, image, audio and video data. Moreover, the technique to merge these different kinds of data and to further form the multimodal analysis approach for emotion detection is the main problem, and the fusion strategy is summarized in this paper. Finally, some trends about the sentiment cognition and sentiment generation are used, which provides new ways for further research work.

In this paper [8], Mehta, P., & Pandya tend to match on opinion mining and feeling assessment which is an area of web data mining and Machine Learning. It is shown that after effect of examination by utilizing different ML and Lexicon investigation techniques. Outcomes are analyzed to play out an evaluation study and to check the estimation of the present composition. In this way, it will help the future investigators with understanding present beginnings in the configuration of possibility examination.

In this paper [9] Empirical Analysis is used it is an evidence based approach to study the interpretation of information. It works on real-world data, metrics and results rather than theories and concept it performed empirical analysis on the data set and the scholarly impact of the papers are determined.

Finer Semantic Analysis approach is used to detect the conflicts between the Natural language functional requirements by analyzing their finer semantic composition. In this paper [10], semantic metamodel of functional model developed and algorithm for conflict detection is used. This algorithm correctly identified about 94.93% elements of software requirement. This tool can be used to detect the conflict between natural language functional requirements to improve the quality of final requirement.

Two research fields are used in this paper [11], that are Statistical model called LSA model and Active Research are called Ontology it represents domain of knowledge. Rajani, S., & Hanumanthappa used LSA in automated evaluation against human evaluation and it was also used for extracting semantic information from textual information and Ontology technique is used to extract structured information from constructed data sets.

The method to excerpt user qualities from the images posted on social media, explicitly the gender data. The old-fashioned methods rely on manuscript analysis or abuse visual data only from the user profile picture or colors, Merler, M [12] proposed to look at the distribution of semantics in pictures coming from the whole feed of a person to estimate their gender. In order to compute such semantic distribution, a trained models from existing visual taxonomies to recognize objects, scenes and activities are used and applied on the images in each user's feed. They conducted experiments on a set of ten thousand twitter users and their collection of half a million images revealed that the gender signal can certainly be extracted from the users image feed with 75.6% accuracy. Moreover, the combination of visual cues resulted almost as strong as textual analysis in predicting the gender, while providing corresponding information that can be employed to further improve the gender prediction accuracy to 88% and when combined with textual data as a byproduct of the investigation, Merler, M. is also able to extrapolate the semantic categories of posted pictures mostly correlated to males and females.

In the paper[13] Lipizzi, C. used Twitter data to assess customers initial reactions to the launch of two new products by Apple and Samsung by examining the streams generated in a 72 h window around the two events. Lipizzi, C. presented a technique based on conversational analysis to extract concept maps from Twitter streams and to use semantic and topological metrics to compare the conversations. The findings showed that there are significant differences in the structural patterns of these two conversations and that the analysis of these differences can be highly informative about early customer's perceptions and worth judgments associated with the competing products.

Filippov, A. [14] presented a new algorithms to the hybridization of ontological analysis and techniques of knowledge engineering with the approaches of nature language processing for extracting the semantic and emotional component of semi-structured and unstructured text resources. This approach has improved the efficiency to analyze the social media content, specific data and fuzziness of natural language.

III. TECHNIQUES

In this paper we covered the basic understanding about sentiment analysis and techniques that are used for classification. The systematic review has been discovered the various sentiment analysis techniques with their performance parameters. It has been detected that high accuracy of classification depends upon the quality of features and classification algorithm used in it. The datasets used in this analysis is mostly about the reviews, tweets, posts and hashtags and to predict this we have used naive bayes algorithm.

Data was basically gathered from Twitter on various themes, for example, expressions and legislative issues. This twitter corpus contains 29000 tweets.

Preprocessing Dataset

- Insert Dataset
- Remove null values
- Check data category
- Training data and test data
- Scaling of feature

The Naive Bayes process is used in numerous real-life situations such as:

1. Text classification: It is used as a probabilistic learning technique for text classification. Its classifier is one of the most successful known algorithms when it comes to the classification of

text documents, i.e., whether a text document belongs to one or more categories (classes).

2. Spam filtration: It is an example of text classification. This has converted the standard tools to differentiate spam email from real email. Numerous current email services implement Bayesian spam filtering. Many server-side email filters, such as Spam Bayes, Spam Dispatcher, uses this technique.

3. Sentiment Analysis: It can be used to analyze the tone of tweets, comments, and reviews—whether they are negative, positive or neutral.

4. Recommendation System: The Naive Bayes algorithm in combination with collaborative filtering is used to build hybrid recommendation systems which help in predicting if a user would like a given resource or not.

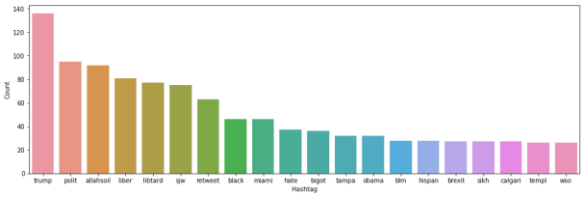

Fig 3. Grapical representation of most used hashtags

Advantages of using naive bayes algorithm-

- It is a comparatively simple process to recognize and form.
- It is quicker to calculate modules using this process than any other grouping processes.
- It is effortlessly skilled using a minor dataset.

IV. RESULT

In this paper we have broadly used the technique of naive bayes to perform the analysis of sentiments on twitter and through this also we have worked on understanding impact of hashtags on tweet sentiment applying the assorted techniques like bag of words, TF-IDF for converting data into features and using these techniques we have acquired an accuracy of 94% in our results.

	Precision	Recall	F1-score	Support
0.0	0.97	0.97	0.97	9367
1.0	0.38	0.42	0.40	465
Accuracy			0.94	9832
Macro average	0.68	0.69	0.68	9832
Weighted average	0.94	0.94	0.94	9832

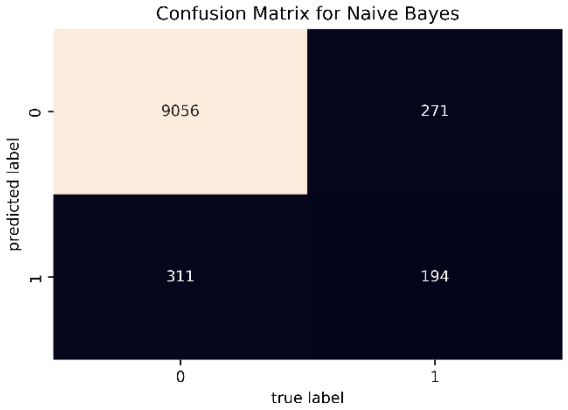

Fig 4. Confusion matrix

V. CONCLUSION

This paper discussed the techniques for sentiment classification and comparison of algorithms experimented by different researchers on different datasets along with performance measures. It is concluded that the Naive Bayes and is the most commonly used algorithm for classification. This

algorithms is mainly used by researches for comparing their proposed work. For sentiment analysis, data is taken from blogs, social media website like Twitter. People freely express their opinion on these media about certain topic, product, and politics. By examining these reviews we have predicted the output with accuracy of 94%. Meanwhile so much of study has been done in the field of sentiment analysis, still it has many challenges. Sometimes people express their opinions in a sarcastic way that is tough to detect. Due to these challenges, sentiment analysis still remains a major area of research.

REFERENCE

[1] J. Ren, S. D. Lee, X. Chen, B. Kao, R. Cheng, and D. Cheung, ―Naive Bayes Classification of Uncertain Data,‖ no. 60703110.

[2] I. Rish, "An Empirical Study of the Naïve Bayes Classifier," no. January 2014.

[3] W. Zhang and F. Gao, ―Procedia Engineering An Improvement to Naive Bayes for Text Classification,‖ vol. 15, pp. 2160–2164, 2011.

[4] Colladon, A. F., D'Angelo, C. A., & Gloor, P. A. (2020). Predicting the future success of scientific publications through social network and semantic analysis. *Scientometrics*, *124*(1), 357-377.

[5] Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, *42*(1), 177-196.

[6] Taghanaki, S. A., Abhishek, K., Cohen, J. P., Cohen-Adad, J., & Hamarneh, G. (2020). Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*, 1-42.

[7] Wang, Y., Rao, Y., & Wu, L. (2017, December). A review of sentiment semantic analysis technology and progress. In *2017 13th International Conference on Computational Intelligence and Security (CIS)* (pp. 452455). IEEE.

[8] Mehta, P., & Pandya, S. A Review On Sentiment Analysis Techniqueologies, Practices And Applications.

[9] Colladon, A. F., D'Angelo, C. A., & Gloor, P. A. (2020). Predicting the future success of scientific publications through social network and semantic analysis. *Scientometrics*, *124*(1), 357-377.

[10] Guo, W., Zhang, L., & Lian, X. (2021). Automatically detecting the conflicts between softwis requirements based on finer semantic analysis. *arXiv preprint arXiv:2103.02255*.

[11] Rajani, S., & Hanumanthappa, M. (2016). Techniques of semantic analysis for natural language processing–a detailed survey. *Int. J. Adv. Res. Comput. Commun. Eng*, *5*(2).

[12] Merler, M., Cao, L., & Smith, J. R. (2015, June). You is what you tweet… pic! gender prediction based on semantic analysis of social media images. In *2015 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE.

[13] Lipizzi, C., Iandoli, L., & Marquez, J. E. R. (2015). Extracting and evaluating conversational patterns in social media: A socio-semantic analysis of customers' reactions to the launch of new products using twitter streams. *International Journal of Information Management*, *35*(4), 490-503.

[14] Filippov, A., Moshkin, V., & Yarushkina, N. (2019, February)