IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Impact Factor: 6.252

Volume 2, Issue 3, June 2022

A Survey on Sentiment Analysis

Abil P Raju[1], Seethal Elias[2]

Software Engineer, AOT Technologies, Idukki, India[1]
Software Engineer, AOT Technologies, Ernakulam, India[2]

Abstract: *The rapid rise of online-based applications, such as forums and blogs, has resulted in comments and updates related to daily activities. Emotional analysis is a process of collecting and analyzing people's ideas, thoughts, and ideas about a variety of topics, products, topics and services. Opinion polls often benefit companies, governments, and individuals by gathering information and making decisions based on ideas. However, emotional analysis and the assessment process face many challenges. These challenges create barriers to accurately defining emotions and determining acceptable emotional differences. Emotional analysis identifies and extracts specific information from the text using language processing and text digging. This article discusses an overview of the strategy for completing this project and for the use of emotional analysis. Then, it evaluates, compares, and investigates common ways to gain a complete understanding of their benefits and disadvantages. Finally, emotional analysis challenges are explored to define future indicators.*

Keywords: Big Data, Machine Learning, Natural Language Processing, Sentiment Analysis, etc.

I. INTRODUCTION

Data is the backbone of the 21st century, and mathematics is a burning machine. Data is everywhere, everywhere assiduity in the form of videos, photos, stats, and textbook. There is no limit to the information as it shows everywhere throughout the macrocosm. As data grows, so does the need for refinement. Every now and then, there are tweets close to Twitter, every nanosecond, nearly 510 comments posted, conditions are edited, and prints are uploaded to Facebook, hourly, Walmart, a series of discount departmental department stores, handles operations more than 1 million customers. Collecting such a large amount of data would be a waste of time, storage space and a problem if it were not used for any meaningful use. Organizations, independent realities, government, political parties and the police, among others, ultimately invest time and plutocrat in extracting data power.

They split the data to understand and interpret application trends, research client gestures, and take financial and planning ideas. The need to sift, sort, organize and distribute this sensitive data systematically leads to the rise of the key word, Big Data. The Big Data is now widely used especially in IT, where it has created colourful job opportunities. The first organizations to adopt it were online and start-up businesses. Businesses like LinkedIn, Google, eBay and Facebook are built around big data in the morning. Big Data is 'like small data', but with a larger size. According to IBM, big data is made by almost everything around us all the time with awesome speed, volume, and versatility. To reward a reasonable amount from Big Data, we need full processing power and chop. Living in the 21st century they are living with a new data perspective known as big data. There are three types of data that we need to consider in Big Data that is organized, structured, and centralized.

A. Structured Data

Structured data can be defined as data with a defined repeating pattern. This pattern makes it easy for any system to filter, read and process data. Processing Built-in Data is much easier and faster than processing data without some recurring patterns. Example: MySQL, Oracle.

DOI: 10.48175/IJARSCT-4911

B. *Unstructured Data*

Random Data is a set of data OR may not have logical or repetitive patterns. It usually contains metadata, i.e. additional data related data, data in various formats such as email, text, audio, video, or images. Example: text, audio, video, opinion, Facebook**.**

C. Built-in Data

Semi-Structured Data, also known as schema-less or self-explanatory, refers to a type of structured data consisting of markers or captions of different objects to distinguish features and generate records and fields in the data provided. Such type of data does not follow the correct configuration of data models as in a related website. Example: CSV but XML and JSON scripts are poorly structured, sensory data, NoSQL data is considered as a central formatstructure.

II. SENTIMENT ANALYSIS

Emotional analysis also known as the concept mine is a field of research that uses people's thoughts, values, beliefs, feelings about products, resources, and people. It focuses on the positive and negative feelings of the user. They are very important in the current situation because, a large number of texts with user ideas are available on the web now. This is a difficult problem to solve because the natural language is not very well organized in nature. [5, 16] Different Levels of Emotional Analysis Different levels of emotional analysis are of three types [17].

A. Document Level

The whole file contains group opinion. The file verifies whether it conveys the positive or negative sentiment.

B. Sentence Level

Verifies that whether a sentence conveys positive, negative or neutral meaning.

C. Entity and Aspect Level

Entity and Aspect Level: Verifies the emotions which are present at each level.

.

III. LITERATURE SURVEY

In [1] this paper explores the problem of classifying texts not by topic, but by feeling as a whole. Naive Bayes is a text editor. Used to assign a given class document. Maximum entropy is another method that has been shown to be effective in many natural language processing systems. Vector support machines are larger margins, rather than powerful, dividing into categories unlike Naive Bayes and Maximum Entropy represented by vector w.

In [2] it automatically collects corporate emotional analysis and objectives. We perform language analysis of aggregated value and provide details of events that are displayed. Using a corpus, it creates an emotional separator that can determine the positive, negative and neutral statements of documents. Using the Twitter API to compile a collection of test posts and create a three-class database. Positive emotions, negative emotions and intentional text set to compile a collection of targeted posts, retrieve text messages from popular twitter accounts. It got 44nwpps to collect a training set of targeted texts, as it exceeds only 140 characters. Then check the frequency distribution of words on the corpus.

In [3] it focuses on a specific domain. Here it begins with the corporation in Malayalam novels, after which the marking (parts-of speech) of the input. Using extracted patterns calculate sentence points. SO-PMI-IR formulas separate input into two desirable or undesirable classes. It appropriately classifies the four categories: happiness, sadness, anger, or neutrality. Simplifying and minimizing the creation of error choruses and pos-tagging tags are done manually.The accuracy of the paper is 63%.

On paper [4] A computer lexicon plays an important role in machine translation while it is a place where all the information about language words is stored for proper practice. Words are a lifeline found in all languages. Every word in a single language has its own meaning and meaning. Syntactic and semantic information about individual words can be incorporated into a pre-programmed repository known as a computer-assisted translation for machine translation. In order to compile a computational dictionary, the first and most important task is to identify the key words or root words in the language. The root word identification plotted in this work is a legal method that without human intervention removes the modified part and finds the root word using morphophonemic rules. The program contains information on 2400 words from the Malayalam corpus to form language data such as root form, their modified forms and grammar section. Presentation is assessed using mathematical actions such as accuracy, memory and f-measure. The criteria obtained for these actions are over 90%.

In [5] this paper shows how to create a state-of-the-art svm state separator. Some are used to get emotional messages such as tweets and sms. Automatically detects message sentiment. That is, determine whether the given message is constructive, positive, or neutral. It will train svm in the training data provided. Each tweet was represented as a feature vector composed of features such as ngrams, ngram characters, pos, hash tags etc .... Some are used to get the feel of the term you have in the message. Automatically detects message sentiment. Automatically detects word feelings in a message. That is, find out if the word in the message conveys a positive, negative, or neutral impression. The same word can express different feelings in different situations.

In [6] a legal way to understand emotions arising from Malayalam reviews. In this paper, the sentence level setting is used. The sentence-level domain is useful for movie websites where the user comments. Emotional analysis uses the rules of denial. It will minimize errors. Here it first collects corpus on movie websites or blogs and newspapers. Using Sandhirules the sentence is re-divided into different tokens. Each word is then matched to a predefined list, and the words are then divided into positive, negative or neutral polarity. After this apply the denial rule to identify the total polarity. Paper accuracy is 85%.

[7] Suggested word resources based on advancing emotions in independent Malayalam reviews. Suggested techniques find a variety of ideas in the text input with the help of a created Hindi Word Net dictionary tool created. Machine learning techniques are used to mark specific situations. This advancement also provides an improved accuracy of 93.6%.

In [8] Speech Components (POS) Marking is the process of transferring to every text display the appropriate POS marker on the outside of it. Consolidation has been the process of identifying and transmitting different types of sentences in sentences. This paper describes a study of the various methods used to mark and classify passages in the Malayalam language. Part of Marking Speech and combining the two most well-known issues in Environmental Language Processing. Tagger can be compared to a translator who reads sentences from a confident language and renders the same sequence of the speech part (POS), which is interesting in the background translation where each word of the sentence comes from. Chunker combines the division of sentences into indistinguishable segments from a very external source of analysis. The tag presented cannot be used in the Indian language. The reasons are: the legal tag will not work because the grammar of Indian languages is very different from western languages and the stochastic mark can be misused. The parts ’marking of the Speech is now a relatively established field. Many methods have been tested.

In [9] provide a customer review section that is used to analyze the most ordered number of tweets and their variations. In this case the first is to process data which is to remove duplicate words and punctuation marks. After pre-processing the data, a set of data is developed and has many different features. Then the feature release removes the adjective from the data set using the unigram model. It also discards prepositions and subsequent words from a sentence. For training and classification use three pre-monitored techniques naïve bayes train classification when the training is completed provides a variety of emotions. Then advanced entropy is widely used in natural language processing an example of that model retrospectiveness associated with a higher level of independent entropy. What follows is a vector equipment to support a supervised learning model with an integrated learning algorithm that analyzes the data used for editing and regression analysis. A set of individual training examples is provided as marked as part of one or two phases. After training and segregation, a semantic analysis based on the data set is used to set each goal associated with the other.

In the [10] mixed method. Used to extract emotions from Malayalam movie reviews. The advanced entropy model is an integrated approach used for marking and specific rules for handling special cases. Such as denials,

DOI: 10.48175/IJARSCT-4911

seals, dilators etc. The upper phase of entropy is a possible phase that is part of the descriptive model phase. Separates inputs without much knowledge. To fit our training data, select high entropy for all models. Marked classes were upgraded to seven from positive, negative and neutral to new classes such as negative, intensifier, dilator and special. The number of positive and negative words is calculated to determine the total polarity. The accuracy is 93.6%.

In [11] work the division of message tweets into positive, negative and neutral. Extremely noisy tweets require prior processing. The precautionary step is to turn the paragraphs into logical words or sentences, and then delete non-English tweets, evoke emotions and delete numbers. It then produces a basic model and pre-processing steps, then learns the positive, negative and intermediate frequencies of the unigram, bigrams, and trigrams in the training set. After the feature domain is active, that is, there are 34 complete features, we count the feature of all the tweets in the message. It was then split into tweet-based features and lexicon-based features. After a preliminary consideration and removal of the feature, the message or tweet will be broken down into positive, negative or neutral emotions.

In [12] perform user feedback analysis using the data mining section. Tweets are collected for training and testing. Then the data is pre-processed, which means clean up data, delete user id, twitter-id, user information, special character, duplicate tweets etc. Then split the tweets into positive, negative, and class neutral by using SentiWordNet 3.0.0 dictionary. Used to set the polarity for each word. If a sentence has zero polarity it means it is neutral, a word that is more positive than negative polarity and then another positive sentence says negative. K-nearest neighbor separator provides better accuracy.

In the empirical analysis program [13] the Malayalam movie review is implemented using a combination of machine learning methods, CRF compliant and SVM compliant rules and applied separately. They concluded that SVM out performs CRF with 91% accuracy.

[14] An effective way of analyzing Hindi tweet emotions: in this paper discuss emotional analysis using Hindi languages. Used by an unregulated lexicon classification method. In this paper choose a subjective lexicon-based method and practice using other bilingual dictionary methods, machine translation and wordnet usage. Here sent Wordnet is built and each dictionary entry is divided into verbs, nouns, adjectives, and extensions. An algorithm based on the proposed subjective lexicon compared to the unigram existence method and the positive and negative words are calculated.

In [15] it is a feature-based analysis of emotions in Malayalam. The main problems that exist in this domain are certain malicious functions only due to the variety, but not the positive function of the characters due to the context in which the user reports. Yet general functioning has been considered in international languages such as English, with few local Malayalam. To create a live movie site if the database includes full movie designs? The emotional names entered by the user; the system will produce the appropriate response according to enteritis. The result can be illustrated in the form of emotional expressions. It lasts three stages. Surveillance contains bold information for registered users as well as movie reviews. Monitoring can enter the movie site and add or remove certain information. Next, user login to the movie site includes a review and post.

The system will produce the same output and display the effect using emotions and icons. Anunregistered user can visit the movie site and see in every movie review. The current system can only test some of the metaphors of emotions. This can be overcome by analyzing additional emotions. Feature-based emotional analysis helps to improve the value of the feature by which the author comments.

[16] contains an analysis of the feelings of the Tamil language. Tamil language expressions are performed with the help of English grammar. Multilingual emotional analysis is divided into two methods which are machine-based translation and a bilingual-based dictionary method. Here are the feelings of review. Tamil language reviews are translated into English with the help of Google translate and translated translations are separated. separations are done using algorithms and techniques etc. Emotional analysis can be done in two ways, namely, the supervised method and the uncontrolled method. The modified method contains polarity separation using the naive Bayes classifier etc ... The unregulated method contains polarity separation using a dictionary of emotions.

DOI: 10.48175/IJARSCT-4911

In [17] it uses emotional analysis in the form of in-depth reading. Here the tweets are collected and processed in advance. Training status and assessment status used for training and database testing. It uses a tensor flow system to create a network. The accuracy is about 67.45% of the train set and the test set is about 52.60%.

E [18] Emotional analysis is the processing of natural language. Take away the reviews of Malayalam movie reviews and classify them as beautiful, negative and neutral. Compared with the English language, Malayalam is a rich language so expressing emotions in Malayalam has some problems. Emotional analysis has certain levels namely, sentence levels for emotional analysis, literary analysis and emotional level analysis. Feeling can be done in three ways such as machine learning and lexicon-based methods. Here a supervised reading machine method is used. The classification is done by specific steps for Web coding of input data, pre-processing (tokens, pos marking), coding of words, extracting straight sentences and editing edits. The proposed system provides about 87.5% accuracy in sentence-level analysis and 90% accuracy in text-grade accuracy.

[19] In the pre-processing of data remove tweet rewriting, text cleaning, handling management etc. then enter a background to use the 3 features of the n-gram word, n-gram character and negative emotion. With the English letter n-gram is better. Supervised learning methods are used to detect hate speech in the Indonesian language. And 10 folded verification is used for testing.

In [20] tweets are directly accessed on the twitter API and create emotional isolation. The author explores people's feelings about a person, product, or genre. By sharing ideas with users about a particular topic they can understand their feelings about it. Here the first collected tweets are specified in the form of hash tags via the twitter API. Verification is done by providing buttons using the RAuth library. After that the certificate is downloaded and a PIN is generated to access the tweets. Using four stages, Category I tweets are collected. Phase II is collected and processed tweets. Tokenization and stemming are performed at this stage. In Phase III the tweets previously considered are compared to the BoW (Word Bag) used for classification as positive, anti-neutral. In the final stage the separated tweets are visualized using a histogram and pie charts.

In [21] this paper explores the emotional analysis through the in-built Python library called Text Blob, to analyze the three twitter forums, the Face book and the news website. Here the author uses the ANN (Artificial Neural Network) to classify tweets. This is a very simple and time-consuming process. ANN is a computer model similar to our biological neural network. Through the twitter API tweets are collected. Separating algorithm tweet naïve bias is used. The feed forward neural network is used to separate data into a train and test using the min-max method where system accuracy is checked. The R setting is used to predict and analyze the outcome. 70-89% accuracy is achieved with a large amount of data used in a very short time.

[22] Emotional analysis or the digging of ideas is a natural language that seeks to capture the emotions of an open mind from a user-generated text. an important topic in the context of movie reviews, product reviews, political discussions etc. User-generated text collected on social media can help machines integrate and capture smart conclusions in a variety of domains. Emotional analysis in Malayalam language is of great importance. Malayalam is a minimalist language and has no standard corpus or emotional dictionary. This activity introduces a machine learning approach to emotional analysis in the Malayalam language using CRF and SVM Learning acquires two levels and the system divides sentences into positive, negative and neutral classes. The work incorporates a large-size design that is described by a corpus as the main function and is followed by teaching a sentence-level separator to analyze emotions.

Paper [23] compares different machine learning with in-depth learning algorithms. In the data collection section tweets are collected using the twitter 140 API. Pre-processing categories have different steps such as noise removal, token making, blocking, punctuation, Word wallet, word deleting etc ... In the download section certain features are identified and detected. Arrangements were made using Naive Bayes, Neural Networks, Decision Trees, Neural Networks, and Random Forrest. The Hybrid model has 83.6% accuracy and 87.1% sensitivity and 79.3% specificity.

DOI: 10.48175/IJARSCT-4911

Table 1:A Summary of Various Works

Year	References	Methodology	Remarks
2012	[3] Mohandas, Neethu,Janardhanan P S Nair, and V. Govindaru	POS tagging	Accuracy of 63%.
2012	[4] Meera Subhash, WilscyM and S.A. Shanavas	Lexicon based	Precision, recall and f-measure of 90%
2014	[6] Nair, D.S.,Jayan, J.P., Rajeev, R.R.,Sherly,E.	Sentence level sentiment analysis	Accuracy of 85%.
2014	[7] Anagha,m, Raveena R Kumar,Sreetha K and P C Reghu Raj	Lexicon based	Accuracy of 93.6%.
2015	[10] Anagha ,M., Raveena R Kumar, Sreetha K and P C Reghu Raj	Maximum entropy	Accuracy is 93.6%.
2015	[13] DeepuS.Nair,JishaP.Jayan,Raheev R.R and Elizabeth Sherly	SVM,CRF	Accuracy 91%.
2017	[18] Ashna M P and Ancy K Sunny	pre-processing-tokenization, stemming	Accuracy: Sentence level-87.5% Document level-90%
2018	[21] SnehPaliwal, Sunil Kumar Khatri and Mayank Sharma,	ANN, R programming	Accuracy 70-89%
2018	[23] Mohammed H. Abd El-Jawad, RaniaHodand Yasser M.k.Omar	Compares deep learning algorithms pre-processing-tokenization, stemming Classification-Naive Bayes, neural network, random forest	Accuracy 83.6 %, sensitivity 87.1% and specificity 79.3%.

IV. CONCLUSION

Daily information is now transmitted in regional languages such as Malayalam, leading to a strong opportunity to analyze this content and determine whether it is acceptable or not. Here in this paper, we analyze the various functions of analyzing emotions and compare the accuracy of the various methods. This paper also reads works in various languages such as Malayalam, English, Tamil etc. Based on research it shows that ANN is much simpler than other machine learning algorithms and is a time-consuming algorithm. ANN is therefore worth analyzing the feelings of multilingualism.

ACKNOWLEDGMENT

We would like convey our sincere gratitude to teachers for their assistance in throughout the procedure. The successful completion of this paper is only because of their proper guidance. We would also like to thank our friends, family, and all others who support us in our work.

REFERENCES

[1] Bo Pang, Lillian Lee and ShivakumarVaithyanathan “Thumbs Up? Sentiment classification using machine learning techniques,” Association forcomputational linguistics, July 2002.

[2] Alexander Pak and Patrick Parouvek,Twitter as a Corpus for Sentiment Analysis and Opinion Mining, In Proceedings of LREC, 2010.

[3] Mohandas, Neethu,Janardhanan P S Nair, and V. Govindaru, “Domain specific sentence level mood extraction from Malayalam text”.Advances in computing and communications (ICACC), 2012 International conference on, IEEE, 2012.

[4] Meera Subhash, Wilscy .M and S.A Shanavas, “A Rule Based Approach for Root Word Identification in Malayalam Language”, International Journal of Computer Science and Information Technology(IJCSIT) Volume:4, June 2012.

[5] Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu “NRC-Canada:Building the state – of – the – Art in Sentiment Analysis of Tweets” Second Joint Conference on Lexical and Computational Semantics, Volume 2, June 2013.

[6] Nair, D.S.,Jayan, J.P., Rajeev, R.R.,Sherly,E. “SentiMa- Sentiment extraction for Malayalam” Advances in computing, Communication and Informatics (ICACCI) 2014 International Conference on IEEE, 2014.

[7] Anagham, Raveena R Kumar, Sreetha K and P C Reghu Raj, “Lexical Resource Based Hybrid Approach for Cross Domain Sentiment Analysis in Malayalam”, International Journal of Engineering Science, Volume:15, December 2014R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, “High-speed digital-to-RF converter,” U.S. Patent 5 668 842, Sept. 16, 1997.

[8] Rinku T S, Merlin Rajan and VarunakshiBhojane, “Various Approaches Used for Tagging and Chunking in Malayalam”, International Journal of Scientific and Engineering Research,Volume:5, May 2014.

[9] Geetika Gautam and Divakar Yadav,” Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis” 7thInternational Conference on Contemporary Computing (IC3), 2014.

[10] Anagha ,M., Raveena R Kumar, Sreetha K and P C ReghuRaj.”A Novel Hybrid Approach on Maximum Entropy Classifier for Sentiment Analysis of M alayalam Movie Reviews” , International Journal of Scientific Research, Volume: 4 June 2015.

[11] Ayush Dalmia, Manish Gupta, Vasudeva Varma “Twitter Sentiment Analysis The good, the bad and the neutral”,Association for Computation Linguistic, June 2015.

[12] Anurag P.Jain and Mr Vijay D. Katkar, “Sentiments Analysis of Twitter Data Using Data Mining”,International Conference on InformationProcessing(ICIP) Vishwakarma Institute of Technology, December 2015.

[13] DeepuS.Nair, JishaP.Jayan, Raheev R.R and Elizabeth Sherly “Sentiment Analysis of Malayalam Film Review Using Machine Learning Techniques”, International Conference on Advancing in Computing, Communication and Informatics (ICACCI) 2015 [14]. Sharma, Y., Mangat, V., and Kaur, “A Practical Approach to Sentiment Analysis of Hindi Tweets”, 1st International Conference on NextGeneration Computing Technologies(NGCT),2015.

[14] Ms Anu P C, MsHeera B M,MsLini K U and Ms.Lilly Raffy Cheerotha, “Aspect Based Sentiment Analysis in Malayalam”, International Journal of Advances in Engineering and Scientific Research, volume:3, December 2016.

[15] R. Thilagavathi and Mrs.K. Krishnakumari, M.E, (PhD), “Tamil English Language Sentiment Analysis System”, International Journal of EngineeringResearch and Technology(IJERT),volume:4,2016.

[16] AdyanMarendraRamadhani and Hong Soon Goo “Twitter Sentiment Analysis Using Deep Learning Methods”, 7th International Annual EngineeringSeminar (InAES), 2017.

[17] Ashna M P and Ancy K Sunny, “Lexicon Based Sentiment Analysis System for Malayalam Language”, International Conference on Computing,Methodologies and Communication (ICCMC) 2017.

[18] IkaAlfina, Rio Mulia, Mohammad Ivan Fanany and YudoEkanata, “Hate Speech Detection in the Indonesian Language: A Data Set and Preliminary Study”, International Conference on Advanced Computer Science and Information System, 2017.

DOI: 10.48175/IJARSCT-4911

[19] PrakruthiV, Sindhu D and Dr. S Anubhama Kumar, “Real Time Sentiment Analysis of Twitter Posts”, 3rd IEEE International Conference onComputation Systems and Information Technology for Sustainable Solutions 2018.

[20] SnehPaliwal, Sunil Kumar Khatri and Mayank Sharma, “Sentiment Analysis and Prediction Using Neural Networks”, Proceedings of the International Conference on Inventive Research in Computing Applications (ICIRCA), 2018.

[21] M. Rahul,R.R Rajeev and S.Shine, “Social Media Sentiment Analysis for Malayalam”,International journal of Computer Sciences and Engineering, Volume:6 2018.

[22] Mohammed H. Abd El-Jawad, Rania Hodhod and Yasser M.k.Omar ,” Sentiment Analysis of SocialMediaNetworks Using Machine Learning”.14thInternational Computer Engineering Conference(ICENCO), 2018

BIOGRAPHY

Abil P Raju received Bachelor of computer applications in 2021 from Santhigiri College of Computer Science Vazhithala ,Idukki affiliated to Mahtma Gandhi University.His research interest is in Deep Learning and Devops. Currently working as a Software Engineer at AOT Technologies,Thiruvananthapuram.

Seethal Elias received Master of Technology in Computer Science and Engineering in 2021 from Mar Athanasius College of Engineering, Kothamangalam affiliated with APJ Abdul Kalam Technological University and received Bachelor of Technology in Computer Science and Engineering from NSS College of Engineering, Palakkad in 2018. Her research interest is in Deep Learning and Data Mining. Currently working as a Software Engineer at AOT Technologies,Thiruvananthapuram.

DOI: 10.48175/IJARSCT-4911