International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Perspective Study on Tweet Sentiment Analysis using Data Mining, Machine Learning and Deep Learning Paradigms

G. Priyadarshini[1], Dr. D. Karthika[2]

[1]Research Scholar, P.K.R.Arts College for Women, Assistant professor, KG College of Arts and Science, India.
[2]Associate Professor and Head Department of computer Science, VET Institute of Arts and Science (Co-Education) College, Erode, TamilNadu, India

Abstract
In recent years there has been an increase in interest in collecting and studying text from social networks, review websites, blogs, forums and other forms of user-generated information. The text offers a vast array of ideas from people of diverse profiles, including education, age and their perspectives, region of residence, on how they see goods and services, policy opinions, etc. The analysis of judgments, responses, and emotions drawn from texts is known as sentiment analysis. The sentiment categorization procedure establishes whether a text is subjective or objective, or whether it provokes both positive and negative responses. The most popular method of classification is based on polarity or orientation for accomplishing tweet sentiment analysis. In this paper, a detailed survey on various algorithms used for performing opinion mining, sentiment analysis, tweet sentiment analysis is discussed in detail. The study shows that text preprocessing, data mining, machine learning algorithm and deep learning paradigms plays a vital role in categorization of people's feeling on a specific topic or a product. In this study, the existing challenges in optimizing the process of tweet sentiment analysis is also discussed and the suggestions for improving is also discussed.

Keywords: Opinion Mining, Tweets, Sentiment Analysis, Machine Learning, Deep Learning, Polarization

Introduction

Although data mining mostly focuses on finding patterns in numerical data, language can also be used to convey information. Contrary to mathematical data, text is usually unorganized and challenging to manage. A subset of text mining is referred known as data mining [1]. The goal of text mining, a young field, is to extract informational value from written or unstructured data. As a result, the term "text mining" describes the act of looking through textual material to discover pertinent patterns. Since text databases are unstructured, using them might be difficult. The practice of extracting keywords, concepts, and other data from various text documents is known as text mining [2]. A common method in opinion mining for identifying sentiments, subjectivities, and sensitive states in online writings is sentiment analysis. The process of product evaluation was finished by organizing the product attributes. Currently,

sentiment polarity assessment is employed in a variety of areas like banking, politics, sports, education, entertainment, etc.

This concentrates on the review of content with a direction, such as opinions or views. A text's subjective or objective nature or whether it stimulates both positive and negative feelings are determined by the sentiment categorization process [3]. There are several significant elements of this classification methodology, such as different procedures, vocations, methodologies, qualities, and application domains. There are many occupations in the classification of sentiment polarity. The three main components of categorization are a class, a level, and an assumption regarding the sentiment sources and objectives.

The distinguishing two-class problem entails categorizing emotions as positive or negative. Sentiment analysis focuses on defining a user's point of view in relation to a given area. Assessment, perception, and even emotional stages are all part of the point of view. The categorization of the polarity of specific text at the levels of characteristics, documents, phrases, and so on is the most important task in sentiment analysis. Applying the classification of polarities, emotional stages such as happy, angry and sad are determined [4].

This classification also identifies the benefits and drawbacks of statements in online reviews, assisting in the more accurate evaluation of products. Agreement detection is another type of binary emotion categorization.

Related Work

More people are using the internet and social media to communicate their views and opinions. As a result, the number of user-generated sentences including sentiment data increased. It's unavoidable to experiment with new ways in order to obtain a better understanding of how people feel and react in various scenarios. Abd et al [5] in their work evaluated the performance of various machine learning and deep learning methods, as well as providing a new hybrid system for sentiment categorization that combines text mining and neural networks. The dataset used in this study contains over one million tweets from five different domains. 75 percent of the dataset was used to train the system, while the remaining 25% was used to test it.

Bing et al. [6] devised a two-step automated method for sorting tweets. To make the task of creating classifiers easier, they used a raucous tutoring set. They classified tweets into subjective and objective categories first and foremost. Subjective tweets are now referred to as "immense" and "negative" tweets.

Zulfadzli et al [7] done a detailed review about sentiment analysis in social media that looked into the methodology, platforms used, and applications. Users have submitted a vast amount of raw material to social media in the form of text, videos, photographs, and audio. The following trusted and credible databases were used to conduct a systematic review of papers published between 2014 and 2019. The publications were evaluated in light of the study's objectives. The findings suggest that the majority of publications used the opinion-lexicon method to analyse text sentiment in social media, extracting data from microblogging sites, mostly Twitter, and using sentiment analysis such as healthcare, business, events and politics.

Dorababu et al [8] contributed sentiment analysis based on the assumption assessment for customers assessment class, which is used to evaluate data in the form of a collection of tweets, where investigations are extremely unstructured and are either high fine or dreadful, or somewhere in the between. For this, initially the dataset is preprocessed dataset, then extracted the adjective from the

dataset that has a couple of significance, referred to as the capacity vector, selected the component vector posting, and then performed device examining using Naive bayes, maximum entropy, and SVM laterally the edge of semantic overview based on word net, which extracts synonyms and similarity for the content.

Luciano et al [9] offered a method for automatically detecting feelings in Twitter messages (tweets) that considers specific features of how tweets are written as well as meta-information about the words that make up these messages. In addition, they use sources of noisy labels as training data. A few sentiment detection websites provided these noisy labels based on twitter data.

Hagen et al [10] categorize the features represented in a tweet as either positive, negative or neutral, we repeat three classification algorithms using different feature sets. The replicated techniques are also merged in an ensemble, with individual classifiers' confidence scores for the three classes averaged and sentiment polarity determined based on these averages.

Deep learning and neural networks have been increasingly important in sentiment analysis in recent years, and they are now widely regarded as the most advanced way for analyzing a variety of languages. Tamil is one of the Indian languages where a state-of-the-art sentiment analysis model is still required. Tamil language presents greater obstacles due to its unique features, grammar structure, and agglutinative nature. To analyse Tamil tweets, Anbukkarasi et al [11] developed a combination of character-based Deep Bidirectional long short-term memory neural networks.

Gangula et al [12] designed a corpora for telugu text and assigning polarities to them is described in this study. Following the establishment of corpora, they trained classifiers to produce accurate classification results. A sentiment classifier is usually trained on data from the same domain that it will be evaluated on. However, there may not be enough data in the same area, and combining data from several sources and domains may aid in the development of a more universal sentiment classifier that can be used across multiple domains. Sentiment data is used from the above corpus from several domains to develop this generalist classifier. For both in-domain and cross-domain categorization, initially examined sentiment analysis models developed with a single data source. Then, using data samples from several areas to construct a sentiment model and validated their performance based on its accuracy of classification.

Hughes et al [13] developed a system for automatically classifying clinical literature at the sentence level to handle complicated text features, a Deep Convolutional Neural Network is used. They used extensive classification of health data to train the model.

Kudo [14] improved the neural network performance by addressing the issue of segmentation vagueness. The regularization method is used to train sub words with a simple regularization. The sampling process is aided by multiple sub word segmentation using the unigram language technique.

Devlin et al [15] anticipated Bidirectional Encoder Representation model, which is a new language representation paradigm. To comprehend the pattern of unlabeled text, the method pretrains bidirectional representation in all layers, both left and right.

Sultana et al. [16] developed a deep learning model which performs sentiment analysis of education data. They used Multilayer perceptron and Support vector machine for training education dataset to predict the sentiment analysis.

Amir Hamzah et al. [17] on their wok used Hidden Markov Model along with POS TAG to determine it is a positive or negative opinion. Automatic detects the orientation of the opinion with its target label.

Wook et al. [18] performed lexicon-based opinion mining which explores the assessment of teaching results. In this work sentiment tendency is analyzed over the student’s feedback which is used for

extracting intensifier words. Teaching assessment is described in terms of positive, neutral or negative opinions.

Kumar Ravi et al. [19] stated in their survey about the various natural language processing techniques, machine learning and sub task performance to achieve the sentiment analysis.

Leary et al. [20] reported in their work about blog mining, which comprised of mining and blog searching. This work analyzed the blog types, its unit and type of opinions to be extracted from blogs.

Montoyo et al. [21] in their work discussed about the problems involved in sentiment and subjectivity analysis and how the machine learning approaches are used to solve the problems in a better manner.

Liu et al. [22] designed a various task conceivable and works based on opinion mining and sentiment analysis. The activities discussed in this work are sentiment classification and subjectivity, sentiment analysis based on aspect, lexicon generation, comparative opinions, summarization of opinion and spam detection opinion and reviews quality are analyzed.

Tsytsarau and Palpanas [23] performed a survey on opinion mining for spam detection, contradiction analysis and aggregation of opinion. Several opining mining approaches are compared with few of the general dataset.

Ali Hasan et al. [24] in their paper designed a hybrid approach which comprised of sentiment analyzing along with machine learning models. This work also performed political view-based sentiment analysis by comparing the support vector machine and naïve bayes.

Kanavos et al. [25] explore an algorithm which handles the emotions from tweets by handling huge volume of dataset to perform sentiment analysis. The author in another work [26] determines the social communities with significant by conveying metric value for users' emotional posts which is collected from twitter profiles.

Alcober et al [27] concentrated on creating a cutting-edge community detection model which will take use of user emotion. The user's tweets are analyzed using Ekman emotional scale, which uses three variants measures and deployed community modularity detection technique.

After more than a year of adjusting to distance learning techniques that are now thought of as the new norm, Mohana et al [28] constructed an opinion mining on the education level of Filipinos. They employed three distinct classification methods to assess opinion mining's accuracy. Final results reveal that deep learning algorithms are most suitable for opinion mining.

Vohra et al [29] designed a finely tuned convolutional network with annotated set of tweets using valence aware dictionary. They adopted fast text embedding model is adopted to training the text corpus dataset. The authors stated that effective extraction of useful text is very challenging to produced best sentiment analysis. work about sentiment analysis on twitter dataset using deep learning and machine learning algorithms. The voting-based ensemble technique is used for binary classification of polarity of the tweet.

Habib et al [30] devised a hybrid deep learning algorithm is used to discover tweet as negative or positive sentence. The long short-term memory improves the classification of either multi-class or binary class categorization by using different word embedding strategies. The work expresses the attitudes, emotions from the opinion of publics.

Limitation in existing algorithms for sentiment analysis and opinion mining

- Class imbalance among the training dataset and testing dataset of tweet and sentiment analysis.
- Overfitting problem when the volume of dataset is huge in opinion mining

- Recognition of accurate user comments is very difficult in case of polarity detection in opinion or sentiment analysis.
- While using conventional Deep learning algorithms the weights are assigned in trial and error basis, then the prediction model produces less accuracy with higher error rate. So, assignment of weight values is very important during classification or prediction process
- The unknown pattern of tweets cannot be categorized by the classification model accurately.

Conclusion

In this paper, the detailed investigation of the existing algorithms used in tweet sentiment or polarity analysis in various domains, such as product, education, books, etc., The most of the classification algorithms focused on dataset with labels. They strongly depend on the training phase of the dataset in sentiment analysis. The text preprocessing, extraction of features and utilizing the significant part of the text are the major challenges to improve the quality of the sentiment analysis. The objective of this survey is to explore the major challenges in text classification based on polarity or opinion mining. The exiting problems can be overcome by focusing on

- Developing classification model with the ability of handling class imbalance
- Construction deep learning models with optimized hyperparameter control
- Designing unsupervised learning model to attain global optimization by fusing optimization algorithms.

References

1. M. H. Abd El-Jawad, R. Hodhod, Y. M. K. Omar, "Sentiment Analysis of Social Media Networks using Machine Learning," *2018 14th International Computer Engineering Conference (ICENCO)*, 2018, pp. 174-176.
2. Yi, Shanshan, Xiaofang Liu, Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers review, Complex & Intelligent Systems 6, no. 3 (2020): 621-634.
3. Asare, A. O., Yap, R., Truong, N. & Sarpong, E. O. The pandemic semesters: Examining public opinion regarding online learning amidst COVID-19. J. Comput. Assist. Learn. 37, 1591–1605, 2021.
4. Kumar Ravi, A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems. 89. 14-46, 2015.
5. M. H. Abd El-Jawad, R. Hodhod, Y. M. K. Omar, "Sentiment Analysis of Social Media Networks using Machine Learning," *2018 14th International Computer Engineering Conference (ICENCO)*, 2018, pp. 174-176.
6. L. Bing, K. C. Chan, C. Ou, Public sentiment analysis in twitter data for prediction of a company's stock price movements, in EBusiness Engineering (ICEBE), 2014 IEEE 11th International Conference on, 2014, pp. 232-239.
7. Zulfadzli Drus, Haliyana Khalid, Sentiment Analysis in Social Media and Its Application: Systematic Literature Review, The Fifth Information Systems International Conference 2019, Procedia Computer Science 161 (2019) 707–714.

8. Dorababu Sudarsa, Siva kumar.P, L. Jagajeevan Rao, Sentiment Analysis for Social Networks Using Machine Learning Techniques, International Journal of Engineering & Technology, 7 (2.32) (2018) 473-476.
9. Luciano Barbosa, Junlan Feng, Robust sentiment detection on Twitter from biased and noisy data, COLING '10: Proceedings of the 23rd International Conference on Computational Linguistics, Pages 36–44, 2010
10. Hagen M., Potthast M., Büchner M., Stein B. (2015) Twitter Sentiment Detection via Ensemble Classification Using Averaged Confidence Scores. In: Hanbury A., Kazai G., Rauber A., Fuhr N. (eds) Advances in Information Retrieval. ECIR 2015. Lecture Notes in Computer Science, vol 9022.
11. S. Anbukkarasi, S. Varadhaganapathy, Analyzing Sentiment in Tamil Tweets using Deep Neural Network, *2020* Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020, pp. 449-453, ICCMC-00084.
12. Gangula Rama Rohit Reddy, Radhika Mamidi, Resource Creation Towards Automated Sentiment Analysis in Telugu (a low resource language) and Integrating Multiple Domain Sources to Enhance Sentiment Prediction, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018, pp 624-634.
13. M. Hughes, I. Li, S. Kotoulas, and T. Suzumura, "Medical Text Classification Using Convolutional Neural Networks," Studies in Health Technology and Informatics, 2017.
14. T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 2018.
15. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
16. J. Sultana, N. Sultana, K. Yadav and F. AlFayez, "Prediction of Sentiment Analysis on Educational Data based on Deep Learning Approach," 2018 21st Saudi Computer Society National Computer Conference (NCC), Riyadh, 2018, pp. 1-5, doi: 10.1109/NCG.2018.8593108.
17. Amir Hamzah, Opinion Mining and Sentiment Analysis Application for Opinion Classification from Education Questionnaire, International Journal of Engineering Research & Technology, Volume 06, Issue 11, 2017
18. Wook, M., Razali, N.A.M., Ramli, S. et al. Opinion mining technique for developing student feedback analysis system using lexicon-based approach (OMFeedback). Educ Inf Technol 25, 2549–2560 (2020). https://doi.org/10.1007/s10639-019-10073-7
19. Kumar Ravi,Vadlamani Ravi, A survey on opinion mining and sentiment analysis: tasks, approaches and applications, Knowledge-Based Systems 89 (2015) 14–46
20. D.E. O'Leary, Blog mining-review and extensions: "From each according to his opinion", Decision Support Systems 51 (2011) 821–830.
21. Montoyo, P. Martínez-Barco, A. Balahur, Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments, Decision Support Systems 53 (2012) 675–679
22. Liu, Sentiment analysis: A multi-faceted problem, IEEE Intelligent Systems 25, no. 3 (2010): 76-80.
23. M. Tsytsarau, T. Palpanas, Survey on mining subjective data on the web, Data Min Knowl Disc (2012) 24:478–514,

24. Ali Hasan, Sana Moin , Ahmad Karim, Shahaboddin Shamshirband, Machine Learning-Based Sentiment Analysis for Twitter Accounts Math and computational applications, 2018, 23, 11; pp 1-15
25. Kanavos, Andreas Nodarakis, Nikolaos Sioutas, Spyros Tsakalidis, Athanasios Tsolis, Dimitrios Tzimas, Giannis, Large Scale Implementations for Twitter Sentiment Classificatio, Algorithms (MDPI). Vol 10 pg 33,2017.
26. Kanavos A, Perikos, I, Hatzilygeroudis, I, Tsakalidis A, Emotional community detection in social networks, Computers & Electrical Engineering Volume 65, Pages 449-460, 2017.
27. G. M. I. Alcober and T. F. Revano, "Twitter Sentiment Analysis towards Online Learning during COVID-19 in the Philippines," 2021 IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), 2021, pp. 1-6.
28. R. S. Mohana, S. Kalaiselvi, K. Kousalya, Twitter based sentiment analysis to predict public emotions using machine learning algorithms, Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 1759-176.
29. Vohra A., Garg R, Deep learning based sentiment analysis of public perception of working from home through tweets. J Intell Inf Syst (2022).
30. Habib Md. Ahsan. (2021). Tweet Sentiment Analysis using Deep Learning Technique. IJITEE (International Journal of Information Technology and Electrical Engineering). 10. 27-36.