IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Document Sentiment Analysis Using Opinion Mining with Machine Learning Techniques

[1]Shivani Dubey 1st, Ajay Kumar Sahu 2nd, Vikas Singhal 3rd

Shan Khan 4th, Sandarbh 5th, Shourav Kumar 6th

Department of Information Technology
Greater Noida Institute of Technology (Engineering Institute), Greater Noida, India

Abstract: Our research aims to shift the focus of sentiment analysis from the product itself to the opinions expressed by users. Existing systems tend to prioritize the product, rather than the sentiments conveyed by users in their reviews. By leveraging natural language processing, text analysis, and computational linguistics techniques, your research intends to semantically analyze and mine user reviews for a given product. The goal is to extract subjective information, identify hidden sentiments, and generate a summarization of the quality and features of the product. This approach will provide a more comprehensive understanding of user opinions and preferences related to the product.

Index Terms **– Opinion Mining, Sentiment Analysis, Machine Learning**

I. INTRODUCTION

Recent years have seen an expansion in the amount of data that is available, or the "data deluge," as a result of more people engaging in electronic activity (such accessing social media online) and as a result of IT being ever more integrated through all gadgets. As an example of the so-called "open data" movement, which is the first of these changes, governments in Access to and use of the data vaults by the general public is growing. Europe and the US have established. Another trend is the massive amount of data that citizens make publicly available through "participatory sensing": everyday people take the initiative to post comments and grievances online and increasingly use technology to collect more data, like images or audio recordings, usually through cell phones. Additionally, sensors being incorporated into commonplace non-ICT objects, such automobiles and urban environments, enabling very quick and automatic data collection of useful information. Finally, both these and government data are now made available to the public, allowing for the continuous flow of new information. Numerous technological advancements at the analytical level aid in making sense of the abundance of data available. We will consider opinion mining in this brief study. The limitations of human attention coupled with the current browser interfaces' basic design Discussion and comments frequently result in low levels of involvement and blazing wars, polarising debate and increasing the likelihood of confrontation. Opinion mining, which deals with subjective statements in contrast to text and pure data mining, was developed to address this problem. This can be seen as a specific evolution of the field of unstructured information extraction (IE), which had previously focused mostly onusing unbiased information, such as statistics on natural disasters or scholarly research. Due to the increase in user-generated material, the availability of public opinion mining techniques to the general public is growing, which increases the potential for applications.

II. SENTIMENT ANALYSIS

In the field of natural language processing (NLP), Sentiment Analysis (SA) has emerged as a significant area of computational research [1], [2]. SA involves mining information related to sentiments or opinions from a group of documents or texts on a specific topic. While some applications focus on sentiments at the document level, other research studies consider opinion-based summarization, emotion or mood extraction, and genre distinctions. SA has gained popularity in various fields such as politics, business, and marketing. It has been used to forecast election outcomes by analyzing sentiments in political forums [3], predict stock market trends by analyzing online sentiments in social media [4], and estimate product sales by examining customer opinions [5]. However, it is important to note that documents assumed to contain sentiments may also include objective information and factual sentences, making it necessary to identify the type and nature of sentences within the SA process. This identification of subjective and objective sentences forms a fundamental part of SA, as subjective sentences are extracted, categorized, and utilized in the analysis. Subjectivity classification is a key task in SA, involving the classification of sentences as objective or subjective. SA typically involves a set of complex processes [6]. The analysis encompasses several tasks, including sentiment classification, subjective analysis, opinion holder extraction, and aspect or object-based extraction. Subjective analysis

entails evaluating a text document or sentence to determine whether it is subjective or objective. Objective documents or sentences are discarded as they are less useful for the SA process. Sentiment classification involves examining the sentiment polarity of the filtered subjective sentences and categorizing them as neutral, negative, or positive. Aspect or object-based extraction is a crucial task in SA, focusing on identifying and extracting the aspects or objects to which opinions are directed. Opinion holder extraction is also important in some cases, as it helps determine the author or source of the opinions. In summary, SA involves a range of tasks, including subjective analysis, sentiment classification, aspect extraction, and opinion holder extraction. These processes collectively contribute to the identification and analysis of sentiments and opinions within textual data

III. RELATED WORK

In our paper, we highlight the importance of consumers having access to reviews and experiences from other consumers who have made similar choices. This helps consumers make informed decisions, avoid mistakes made by others, and clear any confusion about the product or service. You also note that in the current system, leading e-commerce websites primarily focus on the product or its features, with little emphasis on what other people are saying about the product. In contrast, our paper aims to perform opinion mining using sentiment analysis [7]. By semantically analyzing and evaluating product reviews as they are, your research aims to minimize human bias or preference. Instead of comparing products feature-wise, our paper aims to detect hidden sentiments in reviews and combine them with the product's features to provide an overall rating. This approach will assist consumers in the decision-making process during a purchase by providing a comprehensive understanding of both the product's features and the sentiments expressed by other users [8].

IV. METHODS FOR IMPLEMENTATION

Your work is closely related to the research conducted by Hu and Liu in their paper on text mining and summarization [13]. Their research focused on generating feature-based summaries of customer reviews for products sold online. However, there are notable differences in your approach [14]. While traditional text summarization tasks typically focus on summarizing the entire review, your focus is on classifying opinions and features in each sentence of a review. This fine-grained analysis allows for a more detailed understanding of the sentiments expressed throughout the review. To extract the sentiments from the reviews, you employ Part-of-Speech Tagging to identify relevant data sets from the database. In your prototype, you have utilized SentiWordNet for opinion mining. SentiWordNet is a lexical resource that assigns sentiment scores to words based on their semantic similarity to positive and negative terms. By leveraging these techniques, your research aims to provide a comprehensive analysis of user opinions and features in product reviews, enabling a more nuanced understanding of the sentiments expressed by consumers [15].

V. TECHNOLOGY ASPECTS

Among the various approaches available for sentiment analysis (SA), two main groups are widely used. The first group employs the machine learning approach, where multiple techniques are utilized to extract salient features that accurately indicate the polarity of sentiments. This approach requires a manually annotated corpus and constant monitoring of the technique used. The second group adopts a lexicon-based approach, which starts the analysis with words or sentences exhibiting characteristics of semantic polarity. Additionally, there is a combination method, also known as the semi-supervised approach, that combines machine learning with the lexicon-based group [16].

- Machine Learning Approach: In the machine learning approach, two collections of documents are needed: a training collection used by the classifier to differentiate text features, and a test collection used to estimate the classifier's accuracy. Various machine learning algorithms have been developed for text classification into negative or positive classes. Successful approaches include Support Vector Machines (SVM), Naive Bayes (NB), Maximum Entropy (ME), ID3, Centroid Classifier, Winnow Classifier, K-Nearest Neighbor, and Association Rules mining.
- The Naive Bayes (NB) classification method is commonly used for text classification, as it is based on a probabilistic model that estimates the probability of a certain group given a text document input.
- Support Vector Machines (SVM) are proposed as classifiers for pattern recognition between two groups. SVM aims to find the best margin separation of a hyper plane between two groups of data. It has been widely employed for text classification and is considered one of the best methods by many researchers.

Comparisons between machine learning approaches have been conducted to select the best algorithm for sentiment classification. Some studies have shown that Naive Bayes (NB) performs exceptionally well compared to SVM, while others have found that SVM outperforms NB and n-gram models in specific contexts. Other techniques, such as k-means clustering combined with TF-IDF weighting and voting techniques, have also been proposed for sentiment classification. Association Rules mining, a well-researched area, has attracted attention for data exploration approaches, particularly for large datasets such as market basket data in grocery stores. In summary, the machine learning approach, including techniques like Naive Bayes (NB), Support Vector Machines (SVM), and clustering algorithms, has been widely utilized for sentiment classification. These approaches aim to accurately identify sentiment polarity in textual data and have been compared to determine their effectiveness in different contexts. The methodology section outline the plan and method that how the study is conducted. This includes Universe of the study, sample of the study, Data and Sources of Data, study's variables and analytical framework [17].

VI. COMBINATION OF TECHNIQUES

In sentiment analysis (SA), there are different approaches that are commonly used: the machine learning approach, the lexicon-based approach, and a combination of both [18].

A. Machine Learning Approach: The machine learning approach in SA involves using algorithms and techniques to classify texts into positive or negative sentiment categories. Various machine learning algorithms such as Support Vector Machines (SVM), Naive Bayes (NB), Maximum Entropy (ME), and others are commonly employed for sentiment classification. These algorithms utilize training data to learn patterns and features that differentiate between positive and negative sentiment. The performance of these approaches is often evaluated using measures like accuracy, precision, recall, and F1 score.

B. Lexicon-Based Approach: The lexicon-based approach in SA relies on sentiment lexicons or dictionaries that contain words or phrases annotated with their sentiment polarity (positive or negative). This approach assigns sentiment scores to texts based on the presence or frequency of sentiment-bearing words. The semantic orientation (SO) technique is an example of an unsupervised learning approach within the lexicon-based approach. It determines the orientation of a term (positive or negative) based on its proximity to other positive or negative terms in the lexicon. WordNet, a lexical resource, is sometimes used in this approach to identify semantically related terms.

C. Combination Method: The combination method combines machine learning and lexicon-based approaches to improve sentiment classification. It leverages the strengths of both methods to enhance the accuracy and effectiveness of sentiment analysis. Improved versions of classifiers like Naive Bayes and Support Vector Machines may be used, and feature selection techniques such as unigrams and bigrams are applied to improve the classification results.

D. Resources of Sentiment Analysis or Opinion Mining: To perform sentiment analysis, data needs to be collected from various sources such as e-commerce websites like Amazon, social media platforms like Twitter, or using pre-existing sentiment analysis datasets. These resources provide the necessary data for training and testing sentiment analysis models. It's worth noting that different approaches may be more suitable depending on the specific task, dataset, and domain of application. Researchers often compare and evaluate these approaches to select the most effective algorithms for sentiment classification.

VII. DISCUSSION

Complexity compared to topic text classification: Sentiment analysis is considered a specific case of text classification in NLP. However, the process of sentiment classification is more complex than traditional topic text classification. While sentiment analysis deals with a smaller number of classes (positive, negative, neutral), it involves additional challenges due to the nature of the problem. Sentiment analysis goes beyond simple topic classification by aiming to understand and capture the subjective experiences and opinions expressed in text. It requires a deeper understanding of language and context to interpret the subtle nuances, tones, and emotions conveyed by the author. In sentiment analysis, it is essential to consider the sentiment-bearing words and phrases in context, as their meaning can change depending on the surrounding words or the overall sentiment expressed in the text [19]. For example, the word "good" can have a positive sentiment on its own, but in the context of a negative sentence, it may indicate sarcasm or irony. Furthermore, sentiment analysis involves capturing various emotional states, such as joy, anger, sadness, surprise, and fear. These emotions can be expressed through explicit words or implied through figurative language, metaphors, or even emojis. Understanding these nuances requires not only linguistic knowledge but also cultural and contextual awareness. To overcome the challenges of nuanced interpretation, sentiment analysis models often leverage machine learning techniques such as natural language processing (NLP) and deep learning. These models are trained on large datasets to learn patterns and associations between words, phrases, and sentiment labels [20]. They can capture the complexity of language and context to provide more accurate sentiment analysis results.

VIII. CONCLUSION

In conclusion, sentiment analysis, also known as opinion mining, is a crucial area of study for extracting insights from large amounts of unstructured data. It plays a vital role in improving products, services, and overall business management. The existing research in sentiment analysis demonstrates that there is still room for improvement in sentiment classification algorithms and opinion mining techniques. Among the commonly used supervised machine learning approaches for sentiment analysis, Naive Bayes (NB) and Support Vector Machines (SVM) are frequently employed. However, the reviewed approaches indicate that sentiment analyzers are often language-dependent, and there is a lack of a more general and language-independent method. Research and available resources in languages other than English are also limited, highlighting the need for further exploration in this area. The abundance of social media sources, including micro blogs, forums, news platforms, and blogs, provides a vast amount of information regarding people's opinions and sentiments on various subjects. However, utilizing these social media sources for sentiment analysis tasks, particularly in micro-blogging and networking sites, requires more in-depth analysis and research. Challenges in this regard include the difficulty of classification, addressing language generalization issues, and handling negations. Recent developments in natural language processing (NLP) tools have attracted researchers in the field of sentiment analysis, but there is still a need for further improvement. While certain algorithms used in opinion mining or sentiment analysis show promising results, no single method has emerged that can effectively address all the challenges in this domain. Continued research and innovation are necessary to enhance sentiment analysis techniques and overcome existing limitations.

REFERENCES

[1] M. D. Molina-González, E. Martínez-Cámara, M-T Martín-Valdivia and J. M. Perea-Ortega. "Semantic orientation for polarity classification in Spanish reviews". Expert Systems with Applications, , 40(18): 7250-7257, 2013.

[2] P. Chaovalit and L. Zhou "Movie review mining : A comparison between supervised and unsupervised classification approaches". In System Sciences. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on, pp. 112c-112c, 2005. [28] Y. Chen and J. Xie "Online consumer review: Word-of-mouth as a new element of marketing communication mix". Management Science, 54(3): 477-491, 2008. [29] L. Pan. "Sentiment Analysis in Chinese". Brandeis University, 2012.

[3] A. Aue and M. Gamon, "Customizing sentiment classifiers to new domains: A case study". In Proceedings of recent advances in natural language processing (RANLP), pp. 2.1.

[4] A. Esuli and F. Sebastiani."Determining the semantic orientation of terms through gloss classification". In Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 617-624. [32] V. Vapnik "The nature of statistical learning theory". 1995.

[5] Z. Zhang, Q. Ye, Z. Zhang and Y. Li. "Sentiment classification of Internet restaurant reviews written in Cantonese". Expert Systems with Applications, 38(6): 7674-7682, 2011. B. Liu, M. Hu and J. Cheng. "Opinion observer: analyzing and comparing opinions on the web". In Proceedings of the 14th international conference on World Wide Web, pp. 342-351.
[6] J. Brooke, M. Tofiloski and M. Taboada "Cross-Linguistic Sentiment Analysis: From English to Spanish". In RANLP, pp. 50-54.
[7] P. Arora, A. Bakliwal and V. Varma "Hindi subjective lexicon generation using WordNet graph traversal". International Journal of Computational Linguistics and Applications, 3(1): 25-39, 2012.
[8] R. Feldman."Techniques and applications for sentiment analysis". Communications of the ACM, 56(4): 82-89, 2013.
[9] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede Lexiconbased methods for sentiment analysis. Computational linguistics, 37(2): 267-307, 2011.
[10] R. Xia, C. Zong and S. Li. "Ensemble of feature sets and classification algorithms for sentiment classification". Information Sciences, 181(6): 1138-1152, 2011. [12] J. Kamps, M. Marx, R. J. Mokken and M. d. Rijke. "Using WordNet to measure semantic orientations of adjectives". 2004.
[11] Shivani Dubey, Zeba Khanam, Cloud based E Supply Chain Management, International Conference on Supply chain and logistics management (Global Supply chains and Emerging Markets), ICSCLM with collaboration of University of Hull, India Habitat Centre, New Delhi, India, 6-7 December,2013.
[12] Dubey Shivani, Jain Sunayana, An Analytical Framework for Critical Literature Review on Logistics Management via Supply Chain Management, International Conference on Strategy, Innovation & Technology (ICSIT 2014), Ansal University, Gurgaun, March 12- 13, 2014
[13] M. Bautin, L. Vijayarenu and S. Skiena "International Sentiment Analysis for News and Blogs". In ICWSM.
[14] R. Prabowo and M. Thelwall." Sentiment analysis: A combined approach". Journal of Informetrics , 3(2): 143-157,b2009
[15] J. Bollen, H. Mao and X. Zeng "Twitter mood predicts the stock market". Journal of Computational Science, 2(1): 1-8 2011.
[16] E.Baralis, S.Chiusano and P. Garza. (2008)." A Lazy Approach to Associative Classification". IEEE Trans. Knowledge Data Engineering. 20(2): 156-171. [41] R. Agrawal and R. Srikant, (1994) "Fast algorithms for mining association rules". In Proc. of the Int. Conf. on Very Large Databases, pages 487.499,
[17] Shivani Dubey, Mamta Dahiya, Sunayana Jain, Experimental Model of Distributed Service Broker Policy Algorithm in Cloud based Centralized & Distributed Data Center, Journal of Engineering and Applied Science, volume 15, issue 01, 2020.
[18] Shivani Jain, Shivani Dubey, Vikas Singhal, Review of Steganography Techniques for securing Patient Information embedded in Medical Image, International Journal of Scientific Research in Computer Science Applications and Management Studies, Volume 9 Issue 2, ISSN 2319-1953, March 2020.
[19] D. Patil, V. Wadhai, and J. Gokhale, (2010)," Evaluation of Decision Tree Pruning Algorithms for Complexity and Classification Accuracy", International Journal of Computer Applications (0975 – 8887) Volume 11– No.2 [43] Z. Hai, , K. Chang, & J.-J. Kim, 2011. Implicit Feature Identification via Co-Occurrence Association Rule Mining. Dlm. (pnyt.). Computational Linguistics and Intelligent Text Processing, hlm. 393-404. Springer.
[20] W. Y. Kim, J. S.Ryu, , K. I. Kim, & U. M. Kim, 2009. A Method for Opinion Mining of Product Reviews Using Association Rules. Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, hlm. 270- 274.