Sentiment Analysis Approaches: A Systematic Review

Pankaj Kumar Gautam[1,*], Subhadra Shaw[2]

{pankajgautam82@gmail.com, subhadra.shaw@gmail.com}

[1,2]Department of Computer Applications & Information Technology and Sciences, AKS University, Satna, M.P., INDIA

Abstract: The number of consumers who leave online comments has risen dramatically as the number of global Internet presentations continues to rise. Ample data, if correctly explored, should yield important insights. More and more individuals are sharing their ideas and opinions online as a result of the fast rise of social media on the web, forum debates such as reviews, news, comments, and blogs. As a result, this intriguing issue is becoming increasingly significant in business and society. Sentiment analysis is a subcategory of natural language processing and is used to determine the polarity of the user's opinion, which may be positive, negative, or neutral, against the entities or their aspects. SA offers a plethora of current applications in a variety of industries. In the commercial world, it enables organizations to automatically collect client feedback on their products or services. It may be used in politics to infer popular sentiment and reaction to political events, which aids decision-making. This paper explores the various lexicon-based, machine learning (ML), and hybrid techniques used in the field of SA. Hybrid approaches have proved more efficient than any other individual approach.

Keywords: Sentiment Analysis, Lexicon-Based Method, Machine Learning Method, Hybrid Method.

1. Introduction

The modern period has seen an exponential increase in the amount of online data available in various formats [1]. This is due to technological improvements that allow us to be linked 24 hours a day, allowing us to share our experiences and retrieve what others have shared in a short amount of time. This massive amount of online data is a great resource for researchers in a variety of fields [6], [3].

Natural Language Processing (NLP) is one such discipline that aims to teach machines to read text in the same way that people do by using enormous corpora of data. SA is a research field in NLP that is inspired by the inclination to know what others do or feel [6].

Opinion and SA mining is the study of people's feelings, opinions, emotions, attitudes, and evaluations based on documents. Entities in opinion may be organizations, people, products, topics, events, and locations [2]. SA is the study of analyzing people's sentiments, opinions, emotions, appraisals, and attitudes towards entities and their attributes as presented in text [3]. Automatic SA recognizes the writer's emotions without the need for manual methods [11]. Organizations and governments utilize it to gather feedback on their products and services in order to improve them [8] as well as determine whether these expressions are positive, neutral, or negative in nature toward a subject [13]. SA has been used in a variety of fields, including movie recommendations, product reviews, micro blog posts, drama reviews, election result forecasting, and stock market forecasting [10], [4].

THEETAS 2022, April 16-17, Jabalpur, India

DOI 10.4108/eai.16-4-2022.2318164

The rest of the paper is structured as follows: Section 2 Literature Review, Section 3 Sentiment Analysis Approaches Section 4 Comparative Analysis of Various Approaches and Section 5 Conclusion.

2. Literature Review

[1] Performed emotion analysis of users' behavior on reviews from the online dating service (ODS) dataset by combining lexicon-based and semi-supervised ML approaches. The SVM, NB, and KNN classifiers are used. An F1 score of 0.85 proved that a hybrid approach improved classification better than an individual approach.

[2] Explored a "minimally-supervised" approach to build sentiment dictionaries for specialized vocabulary.

[3] Employed deep neural network sequence model, bidirectional long short-term memory with conditional random fields (BiLSTM-CRF) in order to extract target terms from opportunistic statements, and one-dimensional convolutional neural networks (1d-CNNs) for training.

[4] Utilized deep learning to predict the sentiment of Arabic tweets from the Arabic sentiment tweet dataset (ASTD). They have used an ensemble model by combining long-term memory (LSTM) and convolutional neural networks (CNN). This model has a better F1-score of 64.46% than any other individual model.

[5] Focused on a combination of combination of dictionaries and ML models Extreme Gradient Boosting Algorithm (XGBoost) for forecasting the future financial markets behavior of Central Bank speeches

[6] Used an unsupervised hierarchical rule-based technique for extracting aspect terms with a high recall, which they supplemented with pruning algorithms for filtering false positives. The proposed model has a recall rate of 81.9 percent. They have used the Semval 2014 dataset for restaurants and furniture.

[8] Proposed HILATSA (Hybrid Incremental Learning Approach for Arabic Tweet Sentiment Analysis), a hybrid approach by combining lexicons and a ML approach. The classifiers SVM, L2 Logistic Regression, and Recurrent Neural Network (RNN) are employed.

[9] Employed SSentiA (Self-supervised Sentiment Analyzer), a self-supervised hybrid methodology for sentiment classification from unlabeled data that combines an ML classifier with a lexicon-based method.

[10] Explored hybrid approach by combining linguistic rules method with deep learning method to find sentiment of online hotel reserving servicing.

[11] Focused on a custom long-short-term memory deep learning model that was trained using a new Arabic social media text corpus that was augmented with PoS information and demonstrated good sentiment classification accuracy and F1-score.

[12] Implemented bidirectional encoder representations from transformers (BERT) and deep learning techniques to analyses aspect-based SA of airline reviews on TripAdvisor.

[13] Evaluated deep learning techniques like feed-forward neural networks, recurrent neural networks, convolutional neural networks, and domain-ontology for finding aspect-based SA.

[14] Utilized deep learning methods like Convolutional Neural Networks (CNN), Recurrent Neural Networks Long Short-Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT) with a Bi-LSTM for SA of drug reviews.

[15] Used a KNN supervised classification algorithm to find the sentiment of Twitter reviews subject to COVID-19 vaccine effectiveness.

3. Sentiment Analysis Approach

The three most common approaches to SA are lexicon-based, ML-based, and hybrid approaches (see in Figure 1).

3.1 Lexicon-Based Approach

Lexicon-based approaches use linguistic resources such as sentiment lexicons, which are made up of words and their polarity values. This estimates the document's polarity from the polarity of its words. It has high precision but low recall, in addition to the labor required to develop the lexicons. The lexicon method is used for determining the polarity of sentiment in reviews of vehicles, banks, movies, and travel locations. For assessing sentiments from social media data, VADER, a rule-based approach, was introduced. Lexicon-based approaches have been used in a variety of domains, including Twitter, blogging, product evaluations, and tourism [10].

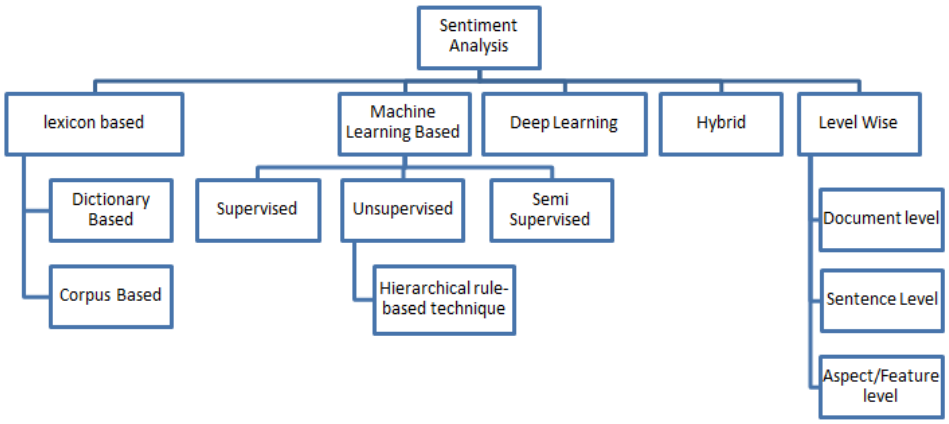

Figure 1: Sentiment Analysis Approaches

Dictionary-based techniques and corpus-based techniques are two classifications of lexicon-based approaches [8].

Dictionary-based techniques start with a predefined dictionary of positive and negative words and then score all of the opinions in the data using word counts or other weighted metrics of word occurrence and frequency. The cost of automatic text analysis is incredibly inexpensive when a lexicon is complete [2]. The dictionary-based approach depends on the algorithm to determine the accuracy of SA. But this approach has a lack of domain specificity. It leads to less efficiency in a particular domain by integrating domain-specific sentiment words into their context or domain.

Corpus-based approach overcomes the limitations of the dictionary-based approach. Although a corpus-based technique has low generalization, it can achieve outstanding performance in a specific domain [8].

3.2 Machine Learning Approach

ML is a subfield of artificial intelligence (AI) that is mainly focused on designing algorithms that can discover patterns and relationships in data [1]. ML-based techniques for sentiment classification are efficient and accurate in situations when training data is available. These strategies, on the other hand, are less useful in situations where there is no training data [2].

ML approaches are classified into supervised, unsupervised, self-supervised, semi-supervised, and hybrid approaches [10].

3.2.1 Supervised Approach

The supervised ML approach is well suited in domains where a large amount of training-labeled data is available to classify the sentiment of text. In either case, when no accurate training data is available, the application of supervised learning approaches introduces inefficiency and potential bias [6].

3.2.2 Unsupervised Approach

The unsupervised ML approach is well suited in domains where unlabeled data is available to train models like Fine-grained SA [11]. Large volumes of labeled training data at the aspect level across diverse domains are always a challenge, necessitating the use of unsupervised models [6].

3.2.3 Semi-supervised Approach

It is the middle ground between unsupervised and supervised learning. It creates a classification model using both labeled and unlabeled data. Self-supervised learning is a type of unsupervised learning in which output labels are automatically created by extracting patterns from input [10].

3.3 Hybrid Approach

Building classifiers such as Support Vector Machine (SVM), Neural Network (NN), and others is essential in an ML-based approach. The classifiers are taught to determine the polarity of documents. The learned models, on the other hand, are domain-specific and necessitate a large number of training datasets. The hybrid approach integrates lexicon-based and ML-based approaches in such a way that the trained model considers the lexicon-based result in its features [9].

There are three levels to SA: the first is the document level, which gives the entire document a single class. The second level is the sentence level, which assigns a class to each sentence. The third level is the feature or aspect level, which identifies the entity's many aspects and computes a separate polarity for each [9], [10]. Aspect level SA is a subfield of SA that operates at a finer grained level to meet the demand of end users who are dissatisfied with an entity's overall sentiment but want to know which features/aspects are of concern and what sentiment is reflected on each of these features/aspects [6].

Table 1. Comparative Analysis of Various Approaches.

Reference	**Method**		**Advantage**	**Limitation**
[1]	Lexicon-Based	Dictionary Based	They don't need labeled data to forecast unobserved events. A fully labeled training set is generally impractical due to the high cost of data labeling, while unlabeled data is comparatively inexpensive. The cost of automatic text analysis is incredibly inexpensive when a dictionary is completed.	When the margin between classes is too small, the explicit lexicon-based strategy frequently fails to differentiate the classes. The lexicon-based system's performance can also be influenced by the dataset's complexity and noise. This approach has a lack of domain specificity. It leads to less efficiency in a particular domain.
[8]		Corpus Based	It can achieve outstanding performance in a specific domain.	A corpus-based technique has low generalization.
[15]	ML Approaches	Supervised	The supervised ML approaches learn implicit patterns from labeled data and hence perform better in determining the polarity of complex scenarios.	The accuracy of supervised ML methods varies by domain and is influenced by parameter adjustment.

[6]		Unsupervised	This is well suited in domains where unlabeled data is available to train models like Fine-grained Sentiment Analysis.	Classification accuracy is low in this approach compared to the supervised approach..
[1]		Semi Supervised	Traditional classifiers take time to label data. Semi-supervised learning has resolved this problem by combining small amounts of labelled data with huge amounts of unlabeled data to build an efficient classifier.	It is complex to build models with this approach.
[8]	Hybrid Approach		It removes the limitations of the component approach and takes advantage of component approaches.	

Conclusion

This study has included various lexicon-based, ML-based, and hybrid methods to find sentiment in various domain-related texts. Each method has its advantages and limitations (see Table 1). The hybrid approach provides a solution to minimize the limitations of those approaches and provide great accuracy.

References

[1] Li H, Chen Q, Zhong Z, Gong R, Han G. E-word of mouth sentiment analysis for user behavior studies. Information Processing & Management. 2022 Jan 1;59(1):102784.

[2] Rice DR, Zorn C. Corpus-based dictionaries for sentiment analysis of specialized vocabularies. Political Science Research and Methods. 2021 Jan;9(1):20-35.

[3] Chen T, Xu R, He Y, Wang X. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. Expert Systems with Applications. 2017 Apr 15;72:221-30.

[4] Heikal M, Torki M, El-Makky N. Sentiment analysis of Arabic tweets using deep learning. Procedia Computer Science. 2018 Jan 1;142:114-22..

[5] Petropoulos A, Siakoulis V. Can central bank speeches predict financial market turbulence? Evidence from an adaptive NLP sentiment index analysis using XGBoost machine learning technique. Central Bank Review. 2021 Dec 1;21(4):141-53.

[6] Venugopalan M, Gupta D. An unsupervised hierarchical rule based model for aspect term extraction augmented with pruning strategies. Procedia Computer Science. 2020 Jan 1;171:22-31.

[7] Nandwani P, Verma R. A review on sentiment analysis and emotion detection from text. Social Network Analysis and Mining. 2021 Dec;11(1):1-9.

[8] Elshakankery K, Ahmed MF. HILATSA: A hybrid Incremental learning approach for Arabic tweets sentiment analysis. Egyptian Informatics Journal. 2019 Nov 1;20(3):163-71.
[9] Sazzed S, Jayarathna S. Ssentia: a self-supervised sentiment analyzer for classification from unlabeled data. Machine Learning with Applications. 2021 Jun 15;4:100026.
[10] Braoudaki A, Kanellou E, Kozanitis C, Fatourou P. Hybrid Data Driven and Rule Based Sentiment Analysis on Greek Text. Procedia Computer Science. 2020 Jan 1;178:234-43.
[11] Nerabie AM, AlKhatib M, Mathew SS, El Barachi M, Oroumchian F. The impact of Arabic part of speech tagging on sentiment analysis: A new corpus and deep learning approach. Procedia Computer Science. 2021 Jan 1;184:148-55.
[12] Chang YC, Ku CH, Le Nguyen DD. Predicting aspect-based sentiment using deep learning and information visualization: The impact of COVID-19 on the airline industry. Information & Management. 2022 Mar 1;59(2):103587.
[13] García-Díaz JA, Cánovas-García M, Valencia-García R. Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in Latin America. Future Generation Computer Systems. 2020 Nov 1;112:641-57.
[14] Colón-Ruiz C, Segura-Bedmar I. Comparing deep learning architectures for sentiment analysis on drug reviews. Journal of Biomedical Informatics. 2020 Oct 1;110:103539.
[15] Shamrat FM, Chakraborty S, Imran MM, Muna JN, Billah MM, Das P, Rahman OM. Sentiment analysis on twitter tweets about COVID-19 vaccines using NLP and supervised KNN classification algorithm. Indones. J. Electr. Eng. Comput. Sci. 2021 Jul;23(1).