International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 06 | June - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

A Review on Data Driven Machine Learning Models for Sentiment Analysis

Archana Parihar[1], Prof. Pankaj Raghuwanshi[2]

Abstract: **Opinion Mining and sentiment analysis of big data has been seen as an active area of research lately. It has a wide range of applications in information systems, including classifying reviews, summarizing review and other real time applications. There are promising possibilities to use sentiment analysis in real time business models. The present work focuses on sentient analysis by classification of tweets from social media (twitter) data. Since the data sets are often extremely large and complex, hence off late, opinion mining and sentiment analysis is executed based on artificial intelligence and machine learning models. This paper presents a systematic review of the existing methods for sentiment analysis of social media data along with their salient contribution.**

Keywords: Data Mining, Opinion Mining, Sentiment Analysis, Machine Learning, Classification Accuracy.

I. Introduction

Sentiment analysis, also known as opinion mining, is a computational technique that involves the use of natural language processing and machine learning algorithms to analyze and determine the sentiment expressed in a piece of text. The primary goal is to discern whether the sentiment conveyed is positive, negative, or neutral. Sentiment analysis finds widespread applications in diverse fields, including social media monitoring, customer feedback analysis, product reviews, and market research. By extracting subjective information from textual data, sentiment analysis enables businesses and organizations to gain valuable insights into public opinion, customer satisfaction, and trends. It plays a crucial role in understanding and responding to user sentiments, helping entities make informed decisions, enhance user experiences, and tailor their strategies to better align with the prevailing sentiments in the online and offline domains. Opinion Mining has emerged as one of the domains of web mining or data mining that has influenced several domains of day to day life. Some of the common examples are

1) Opinion polling
2) Marketing
3) Advertizing
4) Education
5) Politics
6) Finance and Business Predictive Modelling

The sentiment extraction of users from large and complex data sets is however daunting. This is to be ensured that the context (semantics) is to be taken into account prior to reaching conclusions and implicit meaning has to be inferred correctly. Moreover accurate data pre-processing needs to be imposed in order to segregate the useful information from the raw data. Since user sentiments have a critical impact on several parameters and domains, hence sentiment analysis is critically important. While several data sources are available on the internet to be mined, yet a judicious use of web mining is to be done prior to any system design model is to be used. The critical factor is also the feature selection from the raw data to be included in the analysis of the data as a whole.

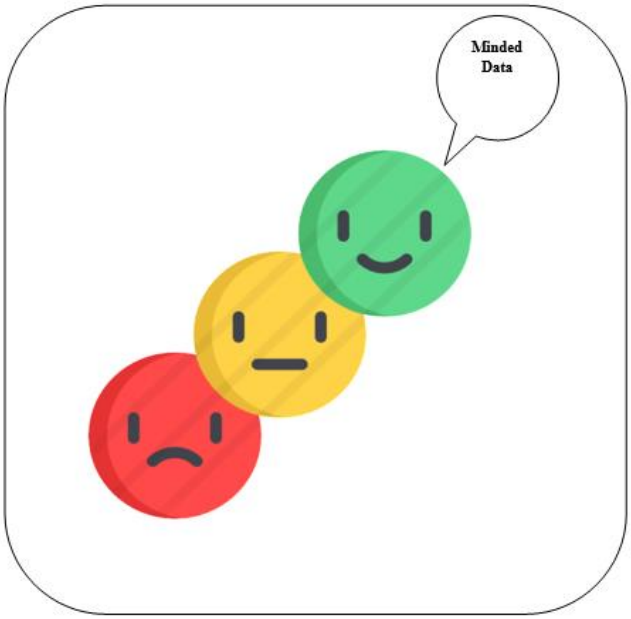

Fig.1 Sentiment Analysis Types

The unstructured text mining approach is often used and the text is to be replaced with suitable tokens or numerical counterparts prior to training any designed mechanism for the classification of the text data [3]. While data as a whole can consist of more than textual data, hence pre-processing of the data is of topmost priority. The automated classification of sentiment based classification can be leveraged in several applications which need an automated mechanism for sentiment classification.

The major challenge in this section is the proper training of the automated system as the training accuracy would yield high classification accuracy later.

II. Literature Review

The literature survey of various scholars related to the review is as follows:-

Zhao et al. [1] proposed a multimodal sentiment analysis method based on the multimodal sentiment analysis method that can obtain more sentimental information sources and help people make better decisions. The experimental results in this paper show that the highest recognition rates of CNN-SVM, 93.5%, respectively.

In [2] author proposed a novel intuitionistic fuzzy inference system (IFIS) for the sentiment analysis. The research paper does the sentiment analysis of using tweets and predicts the personality trait characteristics of the tweeting individual through proposed IFIS. Twitter data was analyzed using Natural Language Processing Toolkit (NLTK) through TextBlob in Google Colaboratory for their subjectivity and polarity to predict the score of their positivity using proposed novel IFIS

Vohra et al. [3] proposed a model uses multiple convolution and max pooling layers, dropout operation, and dense layers with ReLU and sigmoid activations to achieve remarkable results on our dataset. Further, the performance of our model is compared with some standard classifiers like Support Vector Machine (SVM), Naive Bayes, Decision Tree, and Random Forest. From the results, it is observed that on the given dataset, the proposed CNN with FastText word embeddings outperforms other classifiers with an accuracy of 0.925969. As a result of this classification, 54.41% of the tweets are found to show affirmation, 24.50% show a negative disposition, and 21.09% have neutral sentiments towards.

Phan et al. [4] proposed a model which includes the following steps: First, words in sentences are converted vectors using BERT. Second, the contextualized word representations are created based on BiLSTM over word vectors. Third, significant features are extracted and represented using the GCN model with multiple convolutional layers over the contextualized word representations. Finally, the aspect-level sentiments are classified using the CNN model over the feature vectors. Experiments on three benchmark datasets illustrate that our proposed model has improved the performance of the previous context-based GCN methods for ALSA.

Obiedat et al. [5] proposed a hybrid approach by combining the Support Vector Machine (SVM) algorithm with Particle Swarm Optimization (PSO) and different oversampling techniques to handle the imbalanced data problem. SVM is applied as a machine learning classification technique to predict the sentiments of reviews by optimizing the dataset, which contains different reviews of several restaurants in Jordan. Data were collected from Jeeran, a well-known social network for Arabic reviews. A PSO technique is used to optimize the weights of the features.

Vashishtha et al.

[6] proposed MultiLexANFIS which is an adaptive neuro-fuzzy inference system (ANFIS) that incorporates inputs from multiple lexicons to perform sentiment analysis of social media posts. Authors classify tweets into two classes: neutral and non-neutral; the latter class includes both positive and negative polarity. This type of classification will be considered for applications that aim to test the neutrality of content posted by the users in social media platforms. In the proposed model, features are extracted by integrating natural language processing with fuzzy logic; hence, it is able to deal with the fuzziness of natural language in a very efficient and automatic manner.

Saha et al. [7] employed different machine learning algorithms for sentiment analysis. The algorithms such as Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Multilayer Perceptron (MP) on the linguistic features were employed. Authors have determined the precision, recall, F-measure, accuracy, ROC values for each of the classifiers. Among the classifiers Random Forest has outperformed others showing 60.54% correctly classified instance. We believe such sentiment analysis on special category of texts may lead to further investigation in natural language understandings.

Estrada et al. [8] presented a comparison among several sentiment analysis classifiers using three different techniques – machine learning, deep learning, and an evolutionary approach called EvoMSA – for the classification of educational opinions in an Intelligent Learning Environment called ILE-Java. Authors develop two corpora of expressions into the programming languages domain, which reflect the emotional state of students regarding teachers, exams, homework, and academic projects, among others. A corpus called sentiTEXT has polarity

(positive and negative) labels, while a corpus called eduSERE has positive and negative learning-centered emotions (engaged, excited, bored, and frustrated) labels.

Rahat et al. [9] proposed a method in which authors have preprocessed the dataset to convert unstructured airline review into structured review form. After that, we convert structured review into a numerical value. We have to preprocess the data before using it. Stop word removal, @ removal, Hashtag removal, POS tagging, calculating sentiment score have done in preprocessing part. Then an algorithm has been applied to classify the opinion as either it is positive or negative. In this research paper, we will briefly discuss supervised machine learning. Support vector machine as well as Naïve Bayes algorithm and compares their overall accuracy, precession, recall value. The result shows that in the case of airline reviews Support vector machine gave way better result than Naïve Bayes algorithm.

Hasanli et al. [10] developed a roadmap of sentiment analysis of twits in Azerbaijani language. The principles of collecting, cleaning and annotating of twits for Azerbaijani language are described. Machine learning algorithms, such as Linear regression, Naïve Bayes and SVM applied to detect sentiment polarity of text based on bag of word models. Our suggested approach for data processing and classification can be easily adapted and applied to other Turkish language. Achieved results from different machine learning algorithm have been compared and defined optimal parameters for the classification of twits.

III. Existing Methodology

Sentiment Analysis using Machine Learning

Machine learning algorithms are often very helpful in conveying and prioritizing whether a document represents positive, neutral or negative emotions. Machine learning is grouped into two types of extensions such as unsupervised algorithms. The supervised algorithm uses a labeled dataset where each training document is written with positive emotions. While, unsupervised readings include raw data where the text is not written and positive emotions.38 Sentiment analysis is used extensively at 3 categories or levels, namely sentence level, document level, and feature level. Documentary Depression aims to give the entire document or topic good or bad titles. Perceptual deduction considers the minimum of each sentence of a document while the presentation of the index section first identifies the various features of the corpus, and then for each text, the correlation is calculated subjectively to the findings.

Machine Learning Techniques

Sentiment analysis employs various machine learning techniques to analyze and classify the sentiment expressed in textual data. Here are some common machine learning techniques used for sentiment analysis:

Naive Bayes Classifier:

Description: Naive Bayes is a probabilistic classification algorithm that works well for sentiment analysis. It calculates the probability of a document belonging to a particular sentiment class based on the occurrence of words.

Support Vector Machines (SVM):

Description: SVM is a supervised learning algorithm that can classify documents into different sentiment classes. SVM aims to find a hyperplane that separates data points of one sentiment from another, maximizing the margin between classes.

Logistic Regression:

Description: Logistic Regression is a regression analysis adapted for classification tasks. It models the probability of a document belonging to a particular sentiment class and is widely used in binary sentiment analysis.

Random Forest:

Description: Random Forest is an ensemble learning technique that builds multiple decision trees and merges their outputs. It can handle non-linearity and capture complex relationships in the data, making it suitable for sentiment analysis.

Gradient Boosting:

Description: Gradient Boosting is another ensemble method that builds decision trees sequentially, with each tree correcting the errors of the previous ones. It is effective in improving accuracy and handling imbalanced datasets.

Recurrent Neural Networks (RNN):

Description: RNNs are a type of neural network designed for sequence data. They can capture contextual information and dependencies in text, making them suitable for sentiment analysis tasks that involve understanding the sentiment of sentences or paragraphs.

Long Short-Term Memory Networks (LSTM):

Description: LSTMs are a specific type of RNN designed to address the vanishing gradient problem. They excel at capturing long-term dependencies in sequential data, making them effective for sentiment analysis in longer texts.

Convolutional Neural Networks (CNN):

Description: CNNs, commonly used for image recognition, can also be adapted for text classification tasks like sentiment analysis. They use filters to capture local patterns and hierarchical features in textual data.

The choice of machine learning technique depends on factors such as the nature of the data, the scale of the sentiment analysis task, and the available

computational resources. Often, a combination of these techniques or ensemble methods is used to improve overall performance and robustness.

Bayesian Approach

Bayesian regularization is a technique used to introduce Bayesian priors into the learning process, aiding in the regularization of machine learning models. Below is a proposed methodology for incorporating Bayesian regularization into a generic supervised learning framework. Bayesian regularization can be applied to sentiment analysis to improve the robustness and generalization of sentiment classification models, particularly in scenarios where limited labeled data is available

Model Selection

Choose a machine learning model that can benefit from Bayesian regularization. Common models include linear regression, logistic regression, and neural networks. Bayesian regularization is particularly useful when dealing with over fitting.

Bayesian Regularization Approach

The Bayesian Regularization approach introduces an additional penalty factor to optimize the weight updating rule, so as to obtain better regression.

Considering the initial weights as a random variable, w given by:

$$w = [w_1, w_2, \ldots\ldots\ldots w_3]] \quad (1)$$

The penalty factor is defined as ρ=μ/v and is used to update the weights of the networks such that the modified regularized cost function:

$$F(w) = \mu w^T w + v[\frac{1}{n}\sum_{i=1}^{n}(p_i - a_i)^2; \text{ attains a minima} \quad (2)$$

If (π≪v): Network error are generally low.

else if (π≥v): Network errors tend to increase, in which case the weight magnitude should be reduced so as to limit errors (Penalty).

This is done be maximizing the weight Posteriori Probability using the Bayes theorem of Conditional Probability as:

$$P(\langle w \mid X\rangle, \mu, v) \quad (3)$$

Where,

w is the priori probability.

X is the posteriori probability.

The obtained Accuracy is computed as:

$$\text{Accuracy}=(TP+TN)/(TP+TN+FP+FN) \quad (4)$$

Here,

TP, TN, FP and FN denote true positive, true negative, false positive and false negative respectively.

Conclusion:

In conclusion, Sentiment analysis has become a crucial tool for commercial purposes, offering valuable insights into user activities and choices. By employing sentiment analyzers, various methods and algorithms within Natural Language Processing (NLP) are utilized for a comprehensive understanding. This research conducts an extensive review of diverse datasets and research works employing different machine learning techniques for sentiment analysis. Sentiment analysis using machine learning has proven to be a powerful and versatile tool for extracting valuable insights from textual data. The reviewed studies showcased a diverse range of techniques, from traditional methods like Random Forest and Word2Vec to advanced approaches such as Recurrent Neural Networks (RNN). These techniques have demonstrated high accuracy in discerning sentiment across various domains. The future holds promise for continued advancements, particularly in the realm of video sentiment analysis, where the extraction of features from frames and the application of machine learning algorithms are poised to play a pivotal role. Overall, the research underscores the significance of machine learning in unraveling sentiment patterns, contributing to a deeper understanding of user sentiments in diverse applications and domains.

References

[1] Y Zhao, M Mamat, A Aysa, K Ubul, "Multimodal sentiment system and method based on CRNN-SVM", Neural Computing and Applications, Springer, 2023, pp.1-13.

[2] M Dhyani, GS Kushwaha, S Kumar, "A novel intuitionistic fuzzy inference system for sentiment analysis", International Journal of Information Technology, Springer 2022, vol.14., pp. 3193–3200.

[3] A Vohra, R Garg, “Deep learning based sentiment analysis of public perception of working from home through tweets”, Journal of Intelligent Information Systems, Springer 2022, vol.60, pp. 255–274.
[4] H. T. Phan, N. T. Nguyen and D. Hwang, "Aspect-Level Sentiment Analysis Using CNN Over BERT-GCN," in IEEE Access, 2022, vol. 10, pp. 110402-110409.
[5] R. Obiedat R. Qaddoura, A. Al-Zoubi, L. Al-Qaisi, O. Harfoushi, M. Alrefai, H. Faris., "Sentiment Analysis of Customers’ Reviews Using a Hybrid Evolutionary SVM-Based Approach in an Imbalanced Data Distribution," in IEEE Access, vol. 10, pp. 22260-22273, 2022.
[6] S Vashishtha, S Susan, “Neuro-fuzzy network incorporating multiple lexicons for social sentiment analysis”, Applications in computing, Springer 2022, vol.26, pp. 487–4507.
[7] A. Saha, A. A. Marouf and R. Hossain, "Sentiment Analysis from Depression-Related User-Generated Contents from Social Media," 2021 8th Intern, “ational Conference on Computer and Communication Engineering (ICCCE), Kuala Lumpur, Malaysia, 2021, pp. 259-264
[8] MLB Estrada, RZ Cabada, RO Bustillos, “Opinion mining and emotion recognition applied to learning environments”, Journal of Expert Systems, Elsevier 2020, vol. 150., 113265
[9] A. M. Rahat, A. Kahir and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), Moradabad, India, 2019, pp. 266-270.
[10] H. Hasanli and S. Rustamov, "Sentiment Analysis of Azerbaijani twits Using Logistic Regression, Naive Bayes and SVM," 2019 IEEE 13th International Conference on Application of Information and Communication Technologies (AICT), Baku, Azerbaijan, 2019, pp. 1-7.
[11] R. B. Shamantha, S. M. Shetty and P. Rai, "Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Performance," 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019, pp. 21-25.
[12] M. Yasen and S. Tedmori, "Movies Reviews Sentiment Analysis and Classification," 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, 2019, pp. 860-865.
[13] P. Karthika, R. Murugeswari and R. Manoranjithem, "Sentiment Analysis of Social Media Network Using Random Forest Algorithm," 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Tamilnadu, India, 2019, pp. 1-5.
[14] L Zheng, H Wang, S Gao,“ Sentimental feature selection for sentiment analysis of Chinese online reviews”, International journal of machine learning and cybernetics, Springer, 2018, vol.9, pp. 75–84.
[15] R. D. Desai, "Sentiment Analysis of Twitter Data," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 114-117.
[16] A. Bayhaqy, S. Sfenrianto, K. Nainggolan and E. R. Kaburuan, "Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes," 2018 International Conference on Orange Technologies (ICOT), Nusa Dua, Bali, Indonesia, 2018, pp. 1-6.
[17] M. I. Zul, F. Yulia and D. Nurmalasari, "Social Media Sentiment Analysis Using K-Means and Naïve Bayes Algorithm," 2018 2nd International Conference on Electrical Engineering and Informatics (ICon EEI), Batam, Indonesia, 2018, pp. 24-29.
[18] S Liao, J Wang, R Yu, K Sato, Z Cheng, “CNN for situations understanding based on sentiment analysis of twitter data”, Procedia computer science, Elsevier 2017, vol.111, pp. 376-381.
[19] Q. -H. Vo, H. -T. Nguyen, B. Le and M. -L. Nguyen, "Multi-channel LSTM-CNN model for Vietnamese sentiment analysis," 2017 9th International Conference on Knowledge and Systems Engineering (KSE), Hue, Vietnam, 2017, pp. 24-29.
[20] A. Hassan and A. Mahmood, "Deep Learning approach for sentiment analysis of short texts," 2017 3rd International Conference on Control, Automation and Robotics (ICCAR), Nagoya, Japan, 2017, pp. 705-710.