Applying Sentiment Analysis on Children's Stories

Azhar Ahmed Bilal
Department of Computer Engineering
Graduate School of Natural and Applied Sciences
Gazi University
Ankara, Türkiye
aahmed.bilal@gazi.edu.tr
ORCID: 0009-0006-4224-059X

O. Ayhan Erdem
Department of Computer Engineering
Faculty of Technology
Gazi University
Ankara, Türkiye
ayerdem@gazi.edu.tr
ORCID: 0000-0001-7761-1078

Sinan Toklu
Department of Computer Engineering
Faculty of Technology
Gazi University
Ankara, Türkiye
stoklu@gazi.edu.tr
ORCID: 0000-0002-8147-9089

2023 4th International Informatics and Software Engineering Conference (IISEC) | 979-8-3503-1803-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/IISEC59749.2023.10391037

***Abstract*— Today in the world, the rapid and widespread adoption of modern technologies has changed the way that we can access information from news, conversations, and exchange of views about any topics that could attract people's attention. Current communication on the Internet allows people in the community to present and share their ideas in publications and interact with others on different subjects. These data contain a large amount of common knowledge that can be used in scientific and humanitarian research to generalize the benefit to all people everywhere. With the advancement of artificial intelligence and its branches and by using these data, we can solve the various problems that face us in life. Natural language processing is one of the most important issues that have arisen in parallel with the developments of artificial intelligence. Sentiment analysis is regarded as a computational technique that aids in determining and evaluating people's emotions through written documents. Therefore, in this study, we will focus on the sentiment analysis by applying Random Forest and Decision Tree algorithms on the dataset containing children's stories. The results showed that the accuracy rate attained by Random Forest was 98.32% while Decision Tree gave 97.74%.**

Keywords*— *SA (Sentiment Analysis), NLP (Natural Language Processing), AI (Artificial Intelligence), ML (Machine Learning).

I. INTRODUCTION

Nowadays, blogs and social media on the Internet have approved providing an excellent opportunity to express opinions and share experiences on various topics that we encounter in our daily lives [1]. Thus, this information can be used by Internet users to make the best choice about purchasing a service or product or to inquire about something. In fact, opinions matter because they are independent, free from bias, and based on users' experiences. Thus, the quality of a product or service of a company can be improved based on the analysis of the opinions published on the sites. The importance of Artificial Intelligence for data analysis is increasing day by day, this data is becoming the fuel that powers it. The large variety of data is what enables machine learning applications to acquire, and master skills, enabling it to learn and improve its pattern recognition capabilities [2]. So, collecting data from platforms and then analyzing them, helps the researchers to present perfect solutions even to complex problems. Machine Learning and Artificial Intelligence algorithms are utilized to solve a lot of real problems. Researchers in various fields (i.e., psychology, business, computer science, affective computing, artificial intelligence, etc.) have been working on detecting and analyzing sentiment from a piece of text. For example, researchers are able to use data that are mentioned in different posts, tweets and comments to analyze the sentiments of people who discuss various topics [3] [4]. Sentiment Analysis can be defined as considering it as the technique of defining and extracting human feelings through the unstructured text by using Natural Language Processing and Machine Learning techniques. Sentiment analysis is part of text mining, as we mentioned above, the dataset that will be analyzed later can be sourced from the comments from different social media platforms such as Twitter, Facebook, Instagram [5] and other blogs sites. Because social networks have various data that are uploaded from people expressing their opinions or sentiments on any matter, as a result, it can be used as a source for determining sentiment analysis on different subjects. Determining the emotion behind the text is done with the help of NLP (Natural Language Processing) by assigning labels to the texts to put them into categories such as positive, negative, or neutral. Natural language processing (NLP) is a component of Artificial Intelligence (AI) and represents the ability of a computer program to understand human language as it is spoken and written [6]. Natural language processing, which makes progress in a short period, is used for text summarization, machine translation, automatic question answering, sentiment analysis, topic extraction, text mining, asset recognition, parts of speech discrimination, resource extraction and language modeling. One important branch of (AI) Artificial Intelligence is ML (Machine learning) which centers on the using of data and applying different types of algorithms in order to imitate the way that humans learn within improving its accuracy [7].

II. LITERATURE REVIEW

Through the studies and algorithms that have been applied in the field of sentiment analysis of texts, it can be said that the matter still requires high accuracy in order to obtain an excellent result. In [8], in order to separate aspects of the data, classification techniques based on sentiment were used. Also, many traditional methods have suffered from a lack of data and classified dictionaries, and as a result this has led to a decrease in their accuracy. When a feature set has the potential to grow, scalability is a complication that was addressed in the proposed study [9] that performed feature reduction using GA (Genetic Algorithm). This method has greatly reduced the features and the accuracy is somewhat poor. An analysis of the efficiency of the methodology was also conducted using other metrics such as precision, recall, and f-score, in addition to precision. However, the analysis performed on large documents has significant cost and complexity. A person's ability to think and make a decision could be affected by many other aspects of their being altered by the habit of reading books. This is because an individual becomes fully engrossed in the characters and lessons presented in stories [10], particularly in the case of children

979-8-3503-1803-6/23/$31.00 ©2023 IEEE

Authorized licensed use limited to: ULAKBIM UASL - GAZI UNIV. Downloaded on January 23,2024 at 13:55:54 UTC from IEEE Xplore. Restrictions apply.

who experience emotional fluctuations depending on the genre of book they read. Humans experience a range of emotions when they read, including joy, grief, neglect, loneliness, and trauma. The texts that existed in these books and stories have been linked to these emotions [11]. Through the use of NLP (Natural Language Processing), and various classification algorithms that can be used within ML (Machine Learning) techniques, SA identifies, extracts, and characterizes sentiments from the normal text. In order to train and learn the relative associations between different sentiments and emotions, a number of approaches have been used in the SA process. Furthermore, sentiment-based classification methods have been applied to separate the different data aspects [8]. The studied research [12] has attempted to apply dimensionality reduction to documents in the pre-processing phase to reduce the complexity. Sentiment analysis, or SA, is a tool used by NLP (Natural Language Processing) to determine the text's emotional content [13]. It helps identify and classify viewpoints regarding the concept. SA is being used as an effective tool in business to investigate interpretations and consequences for supporting decisions because of the enormous growth in data generated from websites and online applications [14]. Numerous machine learning (ML) has been used to analyze semantics, but more clarification on accuracy is still needed. The naïve bayes algorithm has been used to train data in the examined research [15]. Through software, it has been used to categorize language into positive, negative, and neutral separations. The accuracy of the results is quite high, surpassing the accuracy of the study using Twitter data. In the suggested research [16], the naïve bayes technique has been used for text classification based on coarser assumptions with faster and easier implantation. In this study, we focus on using machine learning algorithms such as Random Forest and Decision Tree as classifications on the dataset, since it provides a higher level of accuracy in predicting outcomes over the decision algorithm.

III. Research Methodology

A. Dataset

This dataset presents 4000 short stories [17] that were categorized based on their emotional content and semantic structure. Many fields are included in the dataset, including the story's URL, length, title, text number, author, and story name. Other fields include text identification, processed text, word count, sentence length, string length, unique words, lexical diversity, and concrete, with other attributes.

B. Pre-processing Data

The Python programming language was used in this study. The current 4000 children's story dataset is loaded and pre-processed during the first stage. By using one of the most NLP libraries, the Natural Language Toolkit (NLTK) which includes packages to help machines comprehend human language and respond to users in the desired manner, the dataset is preprocessed to remove the noise and prepare the data for classification. Before feature extraction, the missing values and the non-standardized feature values are processed using the appropriate scaling values. In the preprocessing operation, data is tokenized, special characters, emoji and links are removed from the dataset, and Vader (Valence Aware Dictionary and Sentiment Reasoner) tool in Python is used in order to label the data. Vader calculates sentiment scores based on the words that are used. It is a rule-based sentiment analyzer where terms are typically classified as positive or negative depending on their semantic orientation [18]. If the polarity score is greater than 0.5 that is considered a positive (happy story), else the polarity score is much less than 0.5 it is considered a negative (sad story).

C. Machine Learning Algorithms

- Random Forest RF: Among the supervised learning methods is the well-known machine learning algorithm Random Forest. It can be applied to ML problems involving both classification and regression. In order to increase the model's robustness and accuracy, the algorithm combines multiple decision trees. Thus, the random forest algorithm for classification tasks will be the main topic of this article.

- Decision Tree: A decision tree algorithm is a type of machine learning algorithm that predicts using a decision tree. It uses a model of decisions and their potential outcomes that resembles a tree. Recursively dividing the data into subsets according to the most important feature at each tree node is how the algorithm operates. One of the greatest algorithms is Random Forest, which reduces overfitting and increases accuracy by combining several decision trees.

TABLE 1. COMPARATIVE PERFORMANCE ASSESSMENT IN TERMS OF ACCURACY METRIC

Classification Models	***Accuracy Metric***	***Precision***	***Recall***	***F1-Score***
Decision Tree	97.74	97	97	97
Random Forest	98.32	98	97	98

IV. Results

This section illustrates the performance analysis of the classification in sentimental analysis on 4000 short stories dataset, with Random Forest and Decision Tree algorithms in terms of different metrics and explicates the comparative analysis section, as showed in Figure1.and Table1.

A. Performance Metrics

1) Accuracy: The term accuracy can be denoted as the model classification rate that is provided through the proportion of correctly classified instances ($Tru_P + Tru_N$) to the sum of instances in the dataset ($Tru_P + Fal_P + Tru_N + Fal_N$) , with the following equation the accuracy range is evaluated.

$$Accuracy = \frac{Tru_N + Tru_P}{Tru_N + Tru_P + Fal_N + Fal_P} \quad (1)$$

Authorized licensed use limited to: ULAKBIM UASL - GAZI UNIV. Downloaded on January 23,2024 at 13:55:54 UTC from IEEE Xplore. Restrictions apply.

2) Precision: The term precision is defined as the degree of covariance of the system, resulted from the correctly identified instances Tru_P to the total number of instances that are correctly classified $(Tru_P + Fal_P)$. It comprises reproducibility and repeatability of the resources. It is measured by equation:

$$Precision = \frac{Tru_P}{Tru_P + Fal_P} \quad (2)$$

In this equation, the variables are defined as., Tru_N-True Negative, Tru_P-True Positive, Fal_N-False Negative, Fal_P-False Positive.

3) F-Measure: F1-score represents weighted harmonic mean value of (Re) recall and (Pr) precision, and it is calculated with the following equation.

$$F - measure = \frac{2*Re*Pr}{Re+Pr} \quad (3)$$

4) Recall: The term recall is the one of the performance metrics, which quantifies the amount of correct positive classification made out of all the positive classification which could have been made. It is computed with the following equation:

$$(Re)Recall = \frac{Tru_p}{Fal_N + Tru_P} \quad (4)$$

B. Comparative analysis

In this study, as shown in the Figure 2. that represents a comparative performance assessment in terms of accuracy metric, different accuracies are obtained for expression classification using the Decision Tree model and Random Forest model. The outcomes of comparative results yield the highest accuracy rate (98.32 %) with Random Forest algorithms outperforming Decision Tree classifiers. This exposes the efficiency of the Random Forest in predicting the highly accurate classifier results. According to Table 1., we remark that the Random Forest algorithm outperforms the Decision Tree algorithm for the four metrics. RF model achieved an accuracy of 98.32 % in comparison with DT (97.74%). The degree of recall reached 97% in both RF and DT. RF has 98 % for the precision and 98 % for the F1-score while DT has 97% in both precision and F1-score.

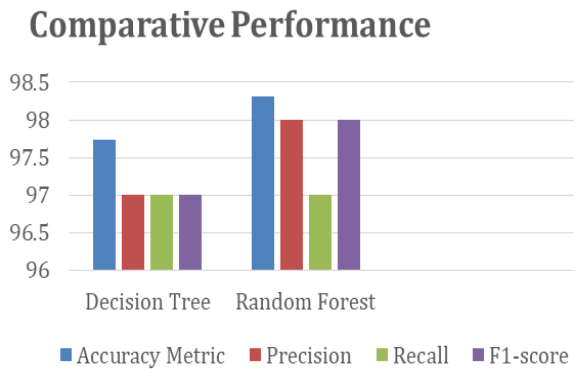


Fig. 1. Comparative performance chart between RF and DT

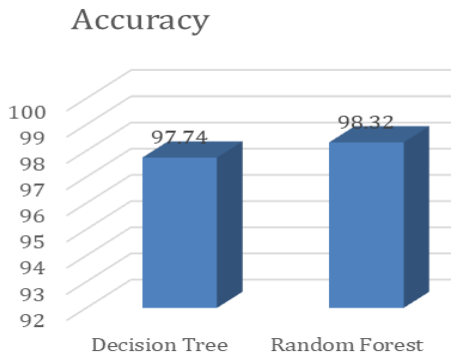


Fig. 2. Comparative performance chart between RF and DT

V. Conclusion

Sentiment classification is an important opinion-mining technique that evaluates viewpoints, opinions, and feelings gleaned from speech, text, and large data sets using various techniques. However, the comparison is unrealistic because this study's work concentrated on just two emotions. The random forest model is excellent, quick, and effective. Due to its high performance and efficiency, it is one of the best techniques and is widely used in many different industries. It is capable of handling categorical, continuous, and binary data. Despite certain drawbacks, random forest is a quick, easy, adaptable, and reliable model.

References

[1] T. A. Rana and Y.-N. Cheah, "Aspect extraction in sentiment analysis: comparative analysis and survey," Artif. Intell. Rev., vol. 46, no. 4, pp. 459–483, 2016, [Online]. Available: https://link.springer.com/article/10.1007/s10462-016-9472-z.

[2] "How Big Data and AI Work Together: Synergies & Benefits." https://www.qlik.com/us/augmented-analytics/big-data-ai (accessed Dec. 29, 2021).

[3] M. H. Abd El-Jawad, R. Hodhod, and Y. M. K. Omar, "Sentiment analysis of social media networks using machine learning," ICENCO 2018 - 14th Int. Comput. Eng. Conf. Secur. Smart Soc., pp. 174–176, Feb. 2019, doi: 10.1109/ICENCO.2018.8636124.

[4] K. P. Dubey and S. Agrawal, "An opinion mining for Indian premier league using machine learning techniques," in 2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), 2019, pp. 1–4.

[5] N. Bahrawi, "Sentiment Analysis Using Random Forest Algorithm-Online Social Media Based," J. Inf. Technol. Its Util., vol. 2, no. 2, pp. 29–33, 2019.

[6] "What is Natural Language Processing? An Introduction to NLP." https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP (accessed Nov. 15, 2023).

[7] "What is Machine Learning? | IBM." https://www.ibm.com/topics/machine-learning (accessed Nov. 15, 2023).

[8] C. Gan, L. Wang, Z. Zhang, and Z. Wang, "Sparse attention based separable dilated convolutional neural network for targeted sentiment analysis," Knowledge-Based Syst., vol. 188, p. 104827, 2020.

[9] F. Iqbal et al., "A hybrid framework for sentiment analysis using genetic algorithm based feature reduction," IEEE Access, vol. 7, pp. 14637–14652, 2019.

[10] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, "An interactive multi-task learning network for end-to-end aspect-based sentiment analysis," arXiv Prepr. arXiv1906.06906, 2019.

[11] P. J. Stone, "Thematic text analysis: New agendas for analyzing text content," Text Anal. Soc. Sci., pp. 35–54, 2020.

Authorized licensed use limited to: ULAKBIM UASL - GAZI UNIV. Downloaded on January 23,2024 at 13:55:54 UTC from IEEE Xplore. Restrictions apply.

[12] N. Li, C.-Y. Chow, and J.-D. Zhang, “SEML: A semi-supervised multi-task learning framework for aspect-based sentiment analysis,” IEEE Access, vol. 8, pp. 189287–189297, 2020.

[13] M. A. Ullah, S. M. Marium, S. A. Begum, and N. S. Dipa, “An algorithm and method for sentiment analysis using the text and emoticon. ICT Express 6 (4): 357–360.” 2020.

[14] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, and I. Awan, “Detection and classification of social media-based extremist affiliations using sentiment analysis techniques,” Human-centric Comput. Inf. Sci., vol. 9, pp. 1–23, 2019.

[15] V. Vangara, S. P. Vangara, and K. Thirupathur, “Opinion mining classification using naive bayes algorithm,” Int. J. Innov. Technol. Explor. Eng., vol. 9, no. 5, pp. 495–498, 2020.

[16] M. Abbas, K. A. Memon, A. A. Jamali, S. Memon, and A. Ahmed, “Multinomial Naive Bayes classification model for sentiment analysis,” IJCSNS Int. J. Comput. Sci. Netw. Secur, vol. 19, no. 3, p. 62, 2019.

[17] “4000 stories with sentiment analysis dataset.” https://brunel.figshare.com/articles/dataset/4000_stories_with_sentiment_analysis_dataset/7712540 (accessed Nov. 16, 2023).

[18] “Sentiment Analysis Using VADER - Analytics Vidhya.” https://www.analyticsvidhya.com/blog/2022/10/sentiment-analysis-using-vader/ (accessed Nov. 16, 2023).

Authorized licensed use limited to: ULAKBIM UASL - GAZI UNIV. Downloaded on January 23,2024 at 13:55:54 UTC from IEEE Xplore. Restrictions apply.