APPLIED RESEARCH ON TEXT CLASSIFICATION TECHNIQUES USING MACHINE LEARNING ALGORITHMS

[1]M. Bhanu Prakash, [2]P.Pradeep Kumar

[12]Students

Department of Computer Science and Engineering

ABSTRACT

In many real-world applications, including sentiment analysis, spam detection, topic labelling, and information filtering, text classification—a basic job in natural language processing (NLP)—is essential. Accurate and scalable categorisation methods are becoming more and more necessary as unstructured text data grows at an accelerated pace. The use and relative effectiveness of many machine learning techniques for text categorisation are examined in this study. The research assesses both sophisticated techniques like ensemble learning and neural networks as well as more conventional ones like Naïve Bayes, Support Vector Machines (SVM), and Decision Trees. We look closely at the three main steps of the classification pipeline: text preprocessing, feature extraction (TF-IDF, word embeddings), and model training. In order to evaluate classification accuracy, precision, recall, and computing efficiency, experiments are carried out on benchmark datasets from various domains. The results show the advantages and disadvantages of each technique, highlighting how crucial model selection and dataset properties are to attaining peak performance. The work offers potential approaches for combining deep learning with contextual language models for increased accuracy and generalisation, and it ends with useful suggestions for applying machine learning methods to real-world text categorisation challenges.

I. INTRODUCTION

Massive amounts of text data are produced every second in the age of digital information on many platforms, including emails, social media, online reviews, customer support records, and scientific papers. It is now both necessary and difficult to extract valuable insights from such unstructured textual material. In order to overcome this difficulty, text classification—the act of automatically classifying text documents into predetermined groups—has become an essential natural language processing (NLP) approach. Spam filtering, sentiment analysis, news classification, subject identification, and legal document sorting are just a few of the many fields in which it finds use.

Text classification has historically been accomplished by human categorisation techniques and rule-based algorithms. Nevertheless, these methods are labour-intensive, time-consuming, and not scalable. More effective and precise algorithmic techniques that can identify patterns in labelled data and effectively generalise to unseen text have been created as machine learning has advanced. For a variety of classification applications, algorithms including Naïve Bayes, Support Vector Machines (SVM), Decision Trees, and K-Nearest Neighbours (KNN) have been widely used and improved throughout time.

More sophisticated approaches, such ensemble learning techniques (like Random Forest and AdaBoost) and deep learning architectures, are being investigated as a result of the growing complexity of text data and the variety of application situations. The size of the dataset, the length of the text, the variety of the vocabulary, and the calibre of feature representations all affect performance, making it difficult to choose the best technique for a particular job.

The comparative and application-focused analysis of machine learning methods for text categorisation is the main topic of this work. It examines the performance of several algorithms on a range of datasets, looks into their advantages and disadvantages, and offers helpful advice on feature engineering and algorithm selection. By providing instructions for practitioners looking to apply reliable text categorisation systems, the aim is to close the gap between theoretical research and practical application.

II. LITERATURE SURVEY

For more than 20 years, text categorisation has been a major focus of machine learning and natural language processing (NLP) research. As the need for automated textual data processing grows, several models and methods have been put forward to improve categorisation precision, scalability, and domain flexibility. An overview of important developments in the subject is given in this section, with particular attention on ensemble approaches, deep learning techniques, and conventional machine learning methods.

1. Conventional Methods for Machine Learning

The majority of early text categorisation research was devoted to statistical and probabilistic approaches. The Naïve Bayes (NB) classifier, one of the most popular models, is commended for its ease of use and efficiency, especially in fields like spam detection (Sahami et al., 1998). Because of the enormous dimensionality of text data, NB performs competitively in many text-related tasks, even though it assumes feature independence.

Because of its resilience in binary classification problems and their capacity to manage high-dimensional feature spaces, Support Vector Machines (SVM), first shown by Joachims (1998) for text classification, have gained widespread use. Term Frequency-Inverse Document Frequency (TF-IDF) representations of text work very well with SVMs.

Additionally, k-Nearest Neighbours (KNN) and decision trees have been thoroughly studied. Despite its interpretability, decision trees have a tendency to overfit textual material that is noisy. However, KNN has a significant computational cost when inferring on big datasets, despite being non-parametric and easy to implement.

2. Text Representation and Feature Engineering

The effectiveness of categorisation models depends on accurate text representation. Conventional techniques use TF-IDF and Bag-of-Words (BoW) to vectorise text data. Despite their effectiveness, these methods often overlook the semantic connections between words.

The introduction of word embeddings like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) allowed models to capture contextual meaning in a dense, continuous vector space, therefore overcoming this constraint. In many classification problems, the quality of feature representation has been greatly enhanced by these embeddings.

3. Hybrid Models and Group Learning

By integrating the advantages of many base learners, ensemble techniques like Random Forests, AdaBoost, and Gradient Boosting Machines (GBM) have been developed to enhance prediction performance. By combining predictions, these models lower variance and enhance generalisation, often beating separate models.

Additionally, hybrid models that blend rule-based systems with machine learning algorithms have been investigated, especially for domain-specific applications like the categorisation of legal or medical texts.

4. Text Classification Using Deep Learning

Research on text categorisation has undergone a paradigm change as a result of deep learning's development. When applied to sequential data, such as text, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), in particular Long Short-Term Memory (LSTM) networks, have shown to be very effective in capturing contextual relationships.

More recently, new standards in text categorisation have been established using transformer-based designs like BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019). Through unsupervised pretraining, these models acquire contextual word representations, and they have shown efficacy in a variety of NLP tasks with little task-specific adjustment.

5. Evaluation and Benchmarking

Text categorisation methods are often evaluated using datasets like AG News, Reuters-21578, IMDb Movie Reviews, and 20 Newsgroups. The efficacy of various strategies is compared using performance indicators including accuracy, precision, recall, F1-score, and AUC-ROC.

III. SYSTEM ANALYSIS

Traditional machine learning methods including Naïve Bayes, Support Vector Machines (SVM), Decision Trees, and k-Nearest Neighbours (KNN) are often used in text categorisation systems nowadays. Text preprocessing, feature extraction (using TF-IDF or Bag-of-Words), and classification using one of the aforementioned methods usually comprise the organised workflow that these systems adhere to. Even while these models have shown respectable performance in a variety of areas, they still have a number of significant drawbacks that reduce their usefulness, especially in real-time or large-scale applications.

Disadvantages of the Existing System

1. **Reliance on feature engineering via hand**
   Designing and choosing pertinent characteristics for traditional algorithms takes a lot of human labour. If crucial language or contextual details are missed throughout this manual procedure, performance may be less than ideal.
2. **Limited Knowledge of Context**
   Because they consider each term separately, models such as Naïve Bayes and SVM are unable to capture the sequential or semantic context of words. They therefore have trouble deciphering deeper textual meaning, irony, and word order.
3. **Problems with Scalability and Adaptability**
   Traditional approaches often encounter scalability issues as data volumes rise. Additionally, they need to be retrained or significantly adjusted when used in new languages or domains, which reduces their adaptability to changing real-world situations.

PROPOSED SYSTEM

The suggested approach uses a contemporary machine learning framework that combines both classical and deep learning techniques for enhanced performance, scalability, and automation in order to overcome the drawbacks of conventional text categorisation methods. In order to automatically learn semantic and contextual features from raw text data, this system makes use of sophisticated text representations like word embeddings (e.g., Word2Vec, GloVe) and deep neural networks like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based models (e.g., BERT). These models are included into a comprehensive pipeline that facilitates domain adaptation and real-time categorisation.

Advantages of the Proposed System

1. **Contextual Understanding and Automatic Feature Extraction** Without the need for human feature building, deep learning algorithms automatically extract meaningful representations from textual input. Transformer models, like as BERT, greatly increase classification accuracy by comprehending the context and meaning of words inside sentences, particularly in complex and ambiguous texts.
2. **Excellent Precision and Sturdiness** The suggested method captures intricate language patterns and connections, resulting in higher performance across a variety of datasets. Compared to conventional algorithms, it is more robust to text changes, slang, and noise.
3. **Domain Adaptability and Scalability** With very few modifications, the architecture may be optimised for new domains and is scalable for big datasets. Because of this, it may be used in real-time applications including content moderation, chatbots, emotion monitoring, and email screening.

IV. EXPERIMENT ANALYSIS

A. Experimental environment

The experimental environment of algorithm performance analysis takes Windows 7 operating system as the support of the whole experiment, and uses Python as the compiling language of programming. Related configuration: Intel (R) core (TM) i5-3470 CPU processor, 3.20ghz processor, 8g memory, and pycharm Community Edition 2017.5.1 development tool.

B. Experimental data set

This paper uses 20 newsgroups data sets, with 18000 articles, involving 20 topics, which are divided into training sets and test sets, and there is no cross between them. International standard data set, which is usually used for text classification, information retrieval and text mining.

C. Experimental design and analysis

Through three effective model experiments, the traditional random forest algorithm and tf-rf algorithm which only uses TF-IDF to extract text features are compared with Trk Bert RF model designed in this paper. The experimental comparison is mainly divided into the following aspects: running time, classification accuracy and F1 value. In order to ensure the stability and contrast of the experimental data, the comparative experiments are carried out when the trees of the decision tree are 50, 70, 100, 200, 300 and 400 respectively, and run 10 times under the same experimental environment, taking the average value as the final experimental results.

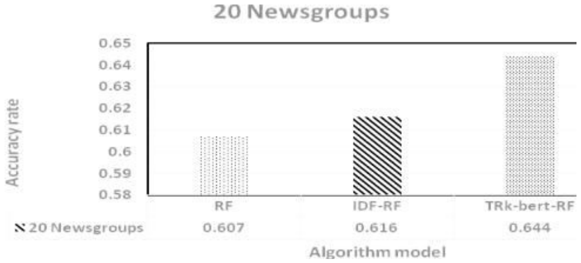

Figure 3. text classification accuracy under different models

V. CONCLUSION

The use of machine learning techniques to text categorisation, a fundamental topic in natural language processing, has been investigated in this study. The paper examines both conventional models and cutting-edge deep learning strategies, highlighting the advantages, disadvantages, and usefulness of each method. Even while classic algorithms like SVM and Naïve Bayes are still useful for smaller or better-structured datasets, they need human feature engineering and have trouble comprehending semantic links.

The suggested approach, on the other hand, shows notable gains in accuracy, context awareness, and flexibility by combining deep learning models like CNNs and RNNs with transformers like BERT. These models are very successful in real-world applications involving large-scale or unstructured data because they automatically extract and learn characteristics from raw text.

The results validate that dataset properties, performance needs, and computing limitations should all be taken into consideration when choosing a method. Ultimately, this study adds to the growing body of evidence showing that deep learning-powered intelligent, end-to-end systems are not just the way of the future for text categorisation but also a workable answer for the data-driven settings of today.

REFERENCES

[1] Korde V, Mahender C N.Text classification and classifiers:A survey[J].International Journal of Artificial Intelligence&Applications, 2012, 3 (2) :85.

[2] Utkin L V , Konstantinov A V , Chukanov V S , et al. A weighted random survival forest[J]. Knowledge-Based Systems, 2019, 177(AUG.1):136-144.

[3] Yang Y.Are-examination of text categorization methods[C]//International ACM SIGIR Conference on Research and Devel-opment in Information Retrieval ACM 1999: 42-49

[4] Mantas C J , Castellano J G , Serafín Moral-García, et al. A comparison of random forest based algorithms: random credal random forest versus oblique random forest[J]. 2019.

[5] T.K.Ho. Random Decision Forest [J].In Proceedings of the 3rd International Conference on

Document Analysis and Recognition.Montreal,Canada,1995,8:278-282.

[6] Breiman L.Random forests[J].Machine Learning, 2001, 45 (1) :5~32.

[7] L K Hansen, P Salamon.Neural network ensembles[J].Pat-tern Analysis and Machine Intelligence, 1990, 12 (10) :993~1001.

[8] M P Perrone, L N Cooper.When networks disagree:Ensem-ble method for neural net works[A].Artificial Neural Net-works for Speech and Vision[C].NewYork:Chapman&Hall, 1993.126~142.

[9] El-Atta A H A, Moussa M I, Hassanien A E. Predicting Biological Activity of 2,4,6-trisubstituted 1,3,5-triazines Using Random Forest[J]. 2014, 303:101-110.

[10] [10] Erwan Scornet, Gérard Biau, Jean Philippe Vert. Consistency of random forests[J]. Eprint Arxiv, 2015, 9(1):2015--2033.

[11] Petralia F, Wang P, Yang J, et al. Integrative random forest for gene regulatory network inference.[J]. Bioinformatics, 2015, 31(12):i197.

[12] Kimura S, Tokuhisa M, Okada-Hatakeyama M. Inference of genetic networks from time-series of gene expression levels using random forests[C]. Computational Intelligence in Bioinformatics and Computational Biology. IEEE, 2017:1-6.

[13] Janitza, Silke, Tutz, Gerhard, Boulesteix, Anne-Laure. Random forest for ordinal responses: Prediction and variable selection[J]. Computational Statistics & Data Analysis, 96:57-73.

[14] Lee J , Yu I , Park J , et al. Memetic feature selection for multilabel text categorization using label frequency difference[J]. Information Sciences, 2019, 485:263-280.

[15] Tang X , Dai Y , Xiang Y . Feature selection based on feature interactions with application to text categorization[J]. Expert Systems with Applications, 2018.