International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 11 | Nov 2022 www.irjet.net p-ISSN: 2395-0072

An in-depth review on News Classification through NLP

Labade Saurabh Ravindra[1], Thorve Rushikesh Vallabh[2], Wavhal Ruturaj Sagar[3], Prof. M. G. Sinalkar[4]

[1,2,3]*Department of Computer Engineering, Jaihind College of Engineering, Pune, Maharashtra, India*
[4]*Asst. Professor, Department of Computer Engineering, Jaihind College of Engineering, Pune, Maharashtra, India*

Abstract - *An ever-increasing number of people in the world depend on online news outlets as their major source of everyday information and current events. As more people discover how useful digital data can be, both the amount of information available and the frequency with which it is accessed are anticipated to grow considerably. With the massive amounts of data being produced by a plethora of publishers, it may be challenging for average consumers to obtain the data they need. Unfortunately, meanwhile, current search engines return so many items that only a small fraction of them are relevant to user requests. As a result, it might be helpful to layer search engines with a classifier, any algorithm created to sort large amounts of data into predetermined categories. There have been considerable inconsistencies that are observed with the realization of an effective news classification approach in the current methodologies which are listed in this survey paper. Thus, to improve this condition there is a need for an effective and useful news classification approach that utilizes Natural Language processing and feature extraction along with Decision Making and fuzzy list. This approach will well defined in the next research on this paradigm.*

Key Words: Natural Language Processing, Feature Extraction, Decision Making, Fuzzy List.

1. INTRODUCTION

As computing power and connectivity have improved over the last several years, so too has the volume of online data. Getting one's news these days mostly comes from the big news portals. A growing amount of news content, unfortunately, poses significant difficulties for news portals. The requirements of modern society cannot be satisfied by using the same text categorization techniques that have been used for decades. As a result, text categorization modeling development has become more important in the area of knowledge discovery in recent times. The speed and accuracy with which the news text categorization technique processes all text input and predicts categorization labels is impressive. As a result, automated classification may provide a cost-effective solution for completing the text categorization job for the media outlet. The study of automated text categorization is becoming more significant in the age of big data.

In today's information age, there are many online resources like Yandex, Bing, and others from which anybody may access a wide range of data. In general, the content of such portals is organized into distinct subcategories for the convenience of viewers. Anyone may quickly and easily get the specific pieces of news and data that most interest them. Included here might be "economic," "educational," "sports," and so on. There are numerous news stories in various categories that are completely unrelated to the topic at hand. In recent years, several studies have been presented for the purpose of news categorization. Researchers in this area have used a wide range of taxonomical approaches to analyses their native tongue.

As more people discover how useful digital information can be, both the volume of information and the frequency with which it is accessed are expected to soar. With the massive amounts of data being produced by a plethora of publishers, it may be challenging for average consumers to get the information they need. Unfortunately, nevertheless, current search engines return so many items that only a small fraction of these are relevant to user queries. As a result, it might be helpful to layer search engines with a classification model, an algorithm created to sort large amounts of data into predetermined categories. Multiple approaches exist that can accurately categorize an English text. The majority of these algorithms are broken down into four sub-categories: text pre-processing, feature extraction, categorization, and effectiveness.

One of the most critical and difficult problems in machine learning techniques is text categorization. Allocating a label from a collection of possible ones to a document corpus autonomously is the task known as text categorization. The significance of this document is determined by the labels that will be applied. For this reason, even for an ordinary human, picking the right collection of labels could be a matter of some ambiguity. The papers may be categorized into one or more categories according to the information we know about them. The documents in this data corpus must only be filed under one of the available categories.

X. Liang et al. present a basic method for identifying CM articles using TF-IDF characteristics, with only middling success [1]. Finally, a graph-based approach is proposed to further improve the identification of content marketing

pieces. To begin, two distinct graph types are developed, each made up of a Sentence Graph and a Word Graph. The authors then offer an innovative technique for determining Sentence Graph edge weights that take into account both semantic similarity and temporal correlations between sentences. Features linked to both graphs and communities may be derived from these two types of graphs. Next, the authors use a dataset that has been manually labeled to train a supervised classifier. The evaluation experiments also reveal that the graph-based features extracted from the graphs can vastly outperform the control group's methods.

The WVDD text similarity measure strategy was introduced by S. Zhou et al. To further increase the rate at which text is clustered, the authors implemented the K-means algorithm using the Spark architecture, which allows for parallel computation. The benefits and viability of the suggested technique are confirmed by F-measure experimental verification, where it is contrasted with the Doc2vec model and the Bag-of-words model [2]. Evidence from experiments shows that WVDD takes into account the structure, word order, and weight settings of Chinese sentences. The suggested technique is more applicable and takes into account semantic relations when the text dataset has a standardized, Chinese sentence component structure and the sentence length is short. Further studies, such as text sentiment analysis, information retrieval, the semantic cognition of artificial intelligence, and so on, might be conducted once this technique yielded more accurate findings for text similarity. When applied to text similarity assessment, a fundamental tool in big data research, the suggested technique can improve the efficiency of information mining.

In this article that analyzes literature, chapter 2 is broken up into an assessment of previous research that is presented in the manner of a reviewed literature, and the third chapter concludes with some suggestions for how more research should be conducted.

2. RELATED WORKS

A hybrid bidirectional recurrent convolutional neural network attention-based model (BRCAN) has been proposed for fine-grained text categorization by J. Zheng et al. This model integrates the Bi-LSTM and CNN efficiently with the aid of the word2vec model and attention mechanism [3]. The proposed model, as is well known, has many benefits, including the following: it captures the contextual information and the semantics of long text by Bi-LSTM to alleviate the problem of information imbalance and save the time-step information; it selects higher-level local features useful for classification from the intermediate sentence representation generated by Bi-LSTM according to the context generated by CNN, and fewer parameters are used to obtain the interaction between hidden layer. Because of this, the proposed model takes the best features from three different models to create a unified representation of a text. The authors compare the proposed model to state-of-the-art classification models that depend on conventional machine learning and deep learning techniques and validate it on multi-topic classification and fine-grained sentiment analysis tasks.

A novel WSD model, proposed by Y. Heo et al., accounts for the increased uncertainty around the senses of some words while also displaying improved prediction accuracy for uncommon senses. Depending on the context of the sample phrase, the model will choose one of several alternative interpretations of the target word. The structure of the Oxford Dictionary inspires this; there, each meaning of a word is recorded and then organized based on its assigned part of speech [4]. Words that have several meanings might benefit from this technique since the scope of possible sense choices is narrowed down. As a bonus, neural language models may now generate unique pieces of context for the target word depending on the segment of speech it is in. The authors also propose a hybrid sense prediction system in which less common word senses are categorized independently from more common ones. This improves prediction accuracy, even if there is significant ambiguity in the sensory environment when working with a small number of training phrases.

W. Zhao et al. present an approach to text categorization that uses only a subset of the available text. In today's online environment, it's usual to encounter financial writings with inconsistent information quality and incomplete textual content. Partial-text-based text categorization may so mimic the incomplete-information environment and is thus more applicable to real-world scenarios. Text classification of subsets of text is more demanding and difficult than text classification of the complete text [5]. The authors propose a neural network, AD-CharCGNN that takes advantage of both charCNN and GRU. In other words, the AD-CharCGNN can glean data from both the temporal and geographical dimensions. Character level data drives the AD-CharCGNN. Whether in Chinese, English, numerals, or other unique characters, messages are surprisingly common. All of the aforementioned characters are supported by the network, and a character-level network may read the whole data set without first having to filter out unnecessary words.

Optimized machine learning and deep learning techniques to identify false news were proposed by D. S. Abdelminaam et al. using COVID-19 and other data sets. Tokenization and stemming were just two parts of a comprehensive sentence analysis that took place in the preparatory phase [6]. There are three datasets in total, with one utilized for training and the other two for testing. TF-IDF and Ngrams are used in the machine learning method to feature analysis, whereas word embedding is used in the deep learning technique. Grid search and Keras tuning are

used to achieve the best possible results from each method. Accuracy, precision, recall, and the F1 measure are all utilized to evaluate the effectiveness of both methods.

D. Alsaleh et al. suggested a CNN-GA mixed classification model for Arabic text. Two big datasets were used to verify the suggested model. The outcomes of using GA-CNN were fantastic. In addition, the model fared well in head-to-head tests against both the baseline and a preexisting technique [7]. Therefore, the improvement in classification accuracy and RMSE for Arabic text achieved by combining CNN with GA was confirmed. Since GA is run throughout the training and validation phase to get the optimal weights, GA-CNN requires more time to calculate than the baseline method.

Two methods, traditional ML and DL, were used by H. Saleh et al. to create a system for detecting bogus news. As the top-performing DL model, the suggested OPCNN-FAKE model has been implemented. Embedding, dropout, convolutional, pooling, flattening, and output layers make up the six-layer suggested OPCNN-FAKE model [8]. Also, it has been fine-tuned with the help of the hyperopt optimization method, wherein the authors tried out a variety of parameter settings for each layer and settled on the optimal combination. Similar to how n-grams with TF-ID have been employed in ML and DL, word embedding feature extraction approaches have also been used.

S. A. Sulaimani et al. [9] propose a new method for multi-class text classification that makes use of Contextual Analysis. Multiple experiments were constructed to evaluate the effectiveness of the proposed method under two conditions: Unbalanced Classes and a Large Number of Classes. Naive Bayes, Support Vector Machines, K-Nearest Neighbors, and Convolutional Neural Networks are used on a Twitter event corpus to compare their performance to other popular classification methods. With an average f1>97.09% and f1>95.27% in the imbalanced classes and the high number of classes experiments, respectively, the results show that the proposed method performs well in classifying brief texts (tweets) into various groups (events). This level of efficiency is generally accepted as adequate for most projects. The interpretability analysis also shows that this technique is straightforward to understand in comparison to the others employed in this research.

A large-scale unsupervised Bangla language dataset (BanglaLM) is proposed for linguistic research by M. Kowsher et al. This work covers investigating the feasibility of fine-tuning a transformer model for a low-resource language like Bangla and training a language model using the biggest dataset yet produced for Bangla [10]. With this research, the mBERT's mixed weights problem for 104 languages, including Bangla (trained on restricted and more structured data only), is fixed. The authors take a look at Bangla-BERT and demonstrate its efficiency across four NLP downstream tasks: sentiment analysis, named entity recognition, binary, and multilevel text classifications. Finally, they demonstrated that the proposed model significantly outperformed mBERT and other non-contextual models in various downstream tasks, like Bangla fasttext, and word2vec.

D. Jung et al. suggested a news-specific model for detecting document inconsistencies using a graph-based approach. By analyzing textual information, GraDID can determine if a given body context is consistent or not. The authors used two tasks in English and Korean to assess the GraDID. Using NELA17, they improve upon previous state-of-the-art results in the area of inconsistent document detection [11]. They also demonstrate how supernode may be utilized for direct categorization due to its ability to effectively collect the whole information. The authors think that method has wide-ranging potential, from assessing the quality of body text to identifying fake news.

A multi-state support vector machine (MSVM) method was discussed by Ravish et al. By employing a distinct model for feature selection and feature extraction, they have been able to fine-tune the model for a wide variety of datasets. The authors employed an improved version of PCA with two phases, dubbed TP-PCA, to extract features. For the Multi-Class Support Vector Machine, they also suggested a feature extraction approach based on Firefly [12]. The suggested technique has been evaluated on 10 various datasets, such as FakeNewsNet, LIAR, ISOT, PolitiFact, etc. Due to a large number of features present in the datasets, feature selection approaches proved more effective than those that employed all of the features during deep learning model training. The feature extraction algorithms performed poorly only on datasets with a low feature count.

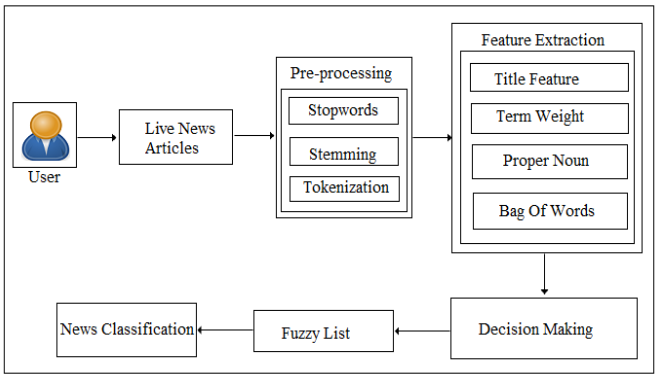

Fig. System Overview Diagram

3. CONCLUSION

Both the amount of data stored and the number of times it is retrieved are projected to increase dramatically as more individuals learn to use the benefits of online

resources. Users may find it difficult to even get the data they require in the ocean of information getting created by a proliferation of sources. However, modern search engine crawlers provide such a massive list of results that even a tiny percentage of them are really appropriate to users' searches. Therefore, it may well be useful to overlay search results with a categorization framework, a technique developed to classify enormous volumes of information into designated groups. There are a number of methods that can successfully classify an English text. This research surveys the current state of the art in news classification techniques and finds that there are many discrepancies that have been discovered with the achievement of a successful news classification methodology. As a result, there is a need for a powerful and practical news categorization strategy that makes use of Natural Language processing, feature extraction, Decision Making, and fuzzy lists to ameliorate the current situation. In subsequent studies using this paradigm, this method will be more clearly described.

REFERENCES

[1] X. Liang, C. Wang, and G. Zhao, "Enhancing Content Marketing Article Detection With Graph Analysis," in IEEE Access, vol. 7, pp. 94869-94881, 2019, DOI: 10.1109/ACCESS.2019.2928094.

[2] S. Zhou, X. Xu, Y. Liu, R. Chang, and Y. Xiao, "Text Similarity Measurement of Semantic Cognition Based on Word Vector Distance Decentralization With Clustering Analysis," in IEEE Access, vol. 7, pp. 107247-107258, 2019, DOI: 10.1109/ACCESS.2019.2932334.

[3] J. Zheng and L. Zheng, "A Hybrid Bidirectional Recurrent Convolutional Neural Network Attention-Based Model for Text Classification," in IEEE Access, vol. 7, pp. 106673-106685, 2019, DOI: 10.1109/ACCESS.2019.2932619.

[4] Y. Heo, S. Kang and J. Seo, "Hybrid Sense Classification Method for Large-Scale Word Sense Disambiguation," in IEEE Access, vol. 8, pp. 27247-27256, 2020, DOI: 10.1109/ACCESS.2020.2970436.

[5] W. Zhao, G. Zhang, G. Yuan, J. Liu, H. Shan, and S. Zhang, "The Study on the Text Classification for Financial News Based on Partial Information," in IEEE Access, vol. 8, pp. 100426-100437, 2020, DOI: 10.1109/ACCESS.2020.2997969.

[6] D. S. Abdelminaam, F. H. Ismail, M. Taha, A. Taha, E. H. Houssein and A. Nabil, "CoAID-DEEP: An Optimized Intelligent Framework for Automated Detecting COVID-19 Misleading Information on Twitter," in IEEE Access, vol. 9, pp. 27840-27867, 2021, DOI: 10.1109/ACCESS.2021.3058066.

[7] D. Alsaleh and S. Larabi-Marie-Sainte, "Arabic Text Classification Using Convolutional Neural Network and Genetic Algorithms," in IEEE Access, vol. 9, pp. 91670-91685, 2021, DOI: 10.1109/ACCESS.2021.3091376.

[8] H. Saleh, A. Alharbi and S. H. Alsamhi, "OPCNN-FAKE: Optimized Convolutional Neural Network for Fake News Detection," in IEEE Access, vol. 9, pp. 129471-129489, 2021, DOI: 10.1109/ACCESS.2021.3112806.

[9] S. A. Sulaimani and A. Starkey, "Short Text Classification Using Contextual Analysis," in IEEE Access, vol. 9, pp. 149619-149629, 2021, DOI: 10.1109/ACCESS.2021.3125768.

[10] M. Kowsher, A. A. Sami, N. J. Prottasha, M. S. Arefin, P. K. Dhar and T. Koshiba, "Bangla-BERT: Transformer-Based Efficient Model for Transfer Learning and Language Understanding," in IEEE Access, vol. 10, pp. 91855-91870, 2022, DOI: 10.1109/ACCESS.2022.3197662.

[11] D. Jung, M. Kim and Y. -S. Cho, "Detecting Documents With Inconsistent Context," in IEEE Access, vol. 10, pp. 98970-98980, 2022, DOI: 10.1109/ACCESS.2022.3204151.

[12] Ravish, R. Katarya, D. Dahiya, and S. Checker, "Fake News Detection System Using Featured-Based Optimized MSVM Classification," in IEEE Access, vol. 10, pp. 113184-113199, 2022, DOI: 10.1109/ACCESS.2022.3216892.