A Comprehensive Analysis of News Classification Using NLP

K. Kishore[1], V. Sujatha[2] , A. Haripriya[3]

[1,2,,3]Asst. Professor, Ashoka Women's Engineering College, Kurnool

Article Info

Page Number: 248 - 252

Publication Issue:

Vol 71 No. 1 (2022)

Article History

Article Received*: 02 February 2022*

Revised*: 10 March 2022*

Accepted*: 25 March 2022*

Publication*: 15 April 2022*

Abstract

A growing number of people rely on online news providers for daily information and current events. As more individuals learn how helpful digital data can be, both its amount and frequency of use are expected to expand. With so many publications producing data, it may be hard for consumers to find what they need. Modern search engines offer so many results that only a small fraction are relevant to user requests. Adding a classifier to search engines may assist classify enormous amounts of data into specified categories. Inconsistencies have been found in the current news classification methods outlined in this survey article. To improve this, a good news categorization approach must use Natural Language Processing, feature extraction, Decision Making, and fuzzy list. The next paradigm research will define this method well.

Index Terms: Natural Language Processing(NLP), News.

1. Introduction

As computing power and connectivity have advanced, so has online data. Today's news comes from large news portals. Increasing news content hurts news portals. Modern culture can't be satisfied by decades-old text categorization algorithms. Text categorization modelling development has proven crucial in knowledge discovery. The speed and accuracy with which news text categorization predicts labels is astonishing. Automated classification may be a cost-effective way for media outlets to classify content. In the age of big data, text categorization research is growing.

In the information age, Yandex, Bing, and others offer a wealth of internet data. Such portals categorise their content into subcategories for ease. Anyone can easily receive their desired news and facts. economic, educational, sports, etc. There are several news stories in unrelated categories. In recent years, various studies have categorised news. Researchers have utilised many taxonomical methodologies to study their native language.

As more individuals learn how helpful digital information is, its volume and frequency of access are projected to grow. With so much data from so many publishers, it may be hard for customers to find what they need. Current search engines return so many results that only a small fraction are relevant. Adding a classification model to search engines may assist filter enormous amounts of data into specified categories. There are multiple ways to appropriately categorise English texts. The majority of these algorithms are classified as text pre-processing, feature extraction, categorization, and effectiveness.

Text categorization is a key problem in machine learning. Text classification is the task of independently labelling a document corpus. The labels on this document determine its meaning. Choosing the correct selection of labels might be ambiguous, even for a human. The papers may be

grouped based on what we know. This data corpus documents must only be filed under one category.

X. Liang et al. present a basic method for identifying CM articles using TF-IDF features [1]. A graph-based approach is proposed to improve content marketing identification. First, a Sentence Graph and a Word Graph are created. The authors offer an innovative technique for determining Sentence Graph edge weights that considers semantic and temporal similarity. These two graphs can yield graph and community-related features. The authors train a supervised classifier using a manually labelled dataset. Experiments show that graph-based features outperform control group methods.

Zhou et al. introduced WVDD text similarity metric. The authors built the K-means technique utilising the concurrent Spark architecture to cluster text faster. The benefits and practicality of the proposed technique are validated by F-measure experimental verification [2]. Experiments reveal that WVDD considers phrase structure, word order, and weight settings. The suggested technique is more suitable when the text dataset has a standardised Chinese sentence component structure and short sentences. When this technique yields more accurate text similarity results, it could be used for text sentiment analysis, information retrieval, AI semantic cognition, etc. Applying the recommended technique to text similarity assessment, a crucial tool in big data research, can improve information mining.

2. Literature Survey

For fine-grained text categorization, J. Zheng et al. presented a hybrid bidirectional recurrent convolutional neural network attention-based model (BRCAN). This model blends Bi-LSTM and CNN with word2vec and an attention mechanism [3]. The proposed model has many benefits, including the following: it captures the contextual information and semantics of long text by Bi-LSTM to alleviate information imbalance and save 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 interac. Because of this, the suggested model combines the best elements of three models to represent a text. The authors compare the proposed model to state-of-the-art classification models that use machine learning and deep learning and validate it on multi-topic classification and fine-grained sentiment analysis tasks.

A new WSD model by Y. Heo et al. accommodates for greater uncertainty in some word senses while improving forecast accuracy for unusual senses. Depending on the sample phrase's context, the model interprets the target word. Each word's meaning in the Oxford Dictionary is categorised by part of speech [4]. Multiple-meaning words may benefit from this strategy because it narrows their meaning options. As a bonus, neural language models can produce unique context for the target word based on its audio segment. The authors also suggest a hybrid sense prediction method that separates less common and more common word senses. This enhances prediction accuracy even if the sensory environment is ambiguous and there are few training phrases.

W. Zhao et al. employ a portion of the text to classify it. In today's online context, financial writings are often inconsistent and incomplete. Partial-text-based text categorization mimics incomplete information and is more suitable to real-world circumstances. Classifying subsets of text is more challenging than the full text [5]. The authors suggest AD-CharCGNN, which uses charCNN and

GRU. The AD-CharCGNN may gather temporal and spatial data. Character-level data drives AD-CharCGNN. In Chinese, English, numbers, or other characters, messages are prevalent. The network supports all the above characters, and a character-level network can read the complete data set without filtering out unneeded words.

Optimized machine learning and deep learning algorithms to identify fake news were proposed by D. S. Abdelminaam et al. In the preparation phase [6], tokenization and stemming were part of a full sentence analysis. Three datasets are used, one for training and two for testing. Machine learning uses TF-IDF and Ngrams for feature analysis, and deep learning uses word embedding. Grid search and Keras tweaking improve each method's results. Accuracy, precision, recall, and F1 are used to evaluate both procedures.

Alsaleh et al. suggested CNN-GA for Arabic text. Two large datasets verified the model. GA-CNN had great results. The model performed well versus the baseline and a known technique [7]. Combining CNN and GA improved Arabic classification accuracy and RMSE. GA-CNN takes longer to calculate than the baseline technique because it's run during training and validation.

H. Saleh et al. employed ML and DL to detect fake news. OPCNN-FAKE is the top-performing DL model. The six-layer OPCNN-FAKE model [8] includes embedding, dropout, convolutional, pooling, flattening, and output layers. The hyperopt optimization approach was used to find the ideal parameter settings for each layer. Word embedding feature extraction is similar to n-grams with TF-ID in ML and DL.

S. A. Sulaimani et al. [9] employ Contextual Analysis to classify multi-class texts. Multiple experiments were conducted to test the approach under two conditions: Unbalanced, several classes. Naive Bayes, SVMs, KNNs, and CNNs are compared on a Twitter event corpus. With an average f1>97.09% and f1>95.27% in the imbalanced classes and high number of classes trials, respectively, the suggested technique classifies brief messages (tweets) well (events). Most initiatives require this level of efficiency. This technique is easy to understand, according to the interpretability analysis.

M. Kowsher et al. offer a large-scale unsupervised Bangla language dataset (BanglaLM). This work investigates the feasibility of fine-tuning a transformer model for a low-resource language like Bangla and trains a language model using Bangla's largest dataset [10]. This research fixes mBERT's mixed weights problem for 104 languages, including Bangla (trained on structured data only). The authors test Bangla-BERT for sentiment analysis, named entity recognition, binary, and multilevel text classifications. In downstream tasks like Bangla fasttext and word2vec, the suggested model beat mBERT and non-contextual models.

D. Jung et al. proposed a news-specific graph-based model for document discrepancies. GraDID analyses text to assess whether a body context is consistent. The authors assessed GraDID using English and Korean tasks. NELA17 improves state-of-the-art inconsistent document detection [11]. They also show how supernode's capacity to capture all information can be used for direct categorization. The authors think this method can measure body text quality and identify bogus news.

Ravish et al. discussed MSVM. By using a separate model for feature selection and extraction, they may fine-tune the model for many datasets. The authors used TP-PCA to extract features. For Multi-Class SVM, they suggested Firefly-based feature extraction [12]. The suggested technique

was tested on 10 datasets, including FakeNewsNet, LIAR, ISOT, and PolitiFact. Due to the enormous number of features in the datasets, feature selection was more successful than using all features during deep learning model training. Only low-feature datasets had bad feature extraction techniques.

3. Conclusion

As more people use internet resources, more data will be saved and retrieved. In the ocean of information created by a profusion of sources, users may have trouble finding the data they need. Modern search engine crawlers produce such a large list of results that only a small fraction are relevant. It may be helpful to overlay search results with a classification framework, a method for grouping large amounts of data. English texts can be classified using several ways. This research reviews the present state of news categorization approaches and finds significant differences with successful news classification. A robust and practical news categorization technique including Natural Language Processing, feature extraction, Decision Making, and fuzzy lists is needed to improve the situation. Future studies employing this paradigm will clarify this method.

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.