JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Hybrid Approach to Text Classification: Combining Textual and Non-Textual Features for Improved Accuracy

Pooja Lodhi, Akhilesh A. Waoo*
AKS University, SATNA, MP

Abstract

Text classification plays an important role in a variety of tasks, including sentiment analysis, topic ranking, and spam detection. Traditional text classification methods often rely on a single model, such as augmentative machine learning (SVMs) or neural networks. However, this approach may struggle to capture the complex nature of natural language statistics. In recent years, hybrid models combining different categories or techniques have emerged as promising approaches to document classification. This article provides a comprehensive overview of previous developments in text classification using hybrid models. We discuss the motivations behind the use of hybrid models, general modeling, and their strengths and weaknesses. Additionally, we examine various hybrid modeling approaches in text classification, including sentiment analysis, topic classification, and text classification. Finally, we discuss research directions and questions that can be asked in the context of document classification using hybrid models. Overall, this review provides important insights into the latest techniques and applications of mixed media in the text category.

Keywords: Hybrid approach, Random Forest Classifier, Logistic Regression, Pipeline, Textual and Non-textual datasets.

Introduction

Text classification, an important task in natural language processing (NLP), involves placing text into categories based on its content. It forms the basis of several NLP techniques, including sentiment analysis, topic classification, spam detection, and text organization. Traditional text classification methods often rely on neural networks such as augmentative machine learning (SVMs), Bayesian classifiers, or predictive networks. However, these methods may encounter difficulties in solving complex natural language problems, which may result in decreased performance.

To overcome these challenges, the hybrid model has emerged as a promising method for document classification. These models combine multiple classes or techniques, combining the strengths of each to improve overall performance. Hybrid models aim to improve classification, robustness, and overall efficiency by combining different processes. They often combine traditional machine learning algorithms with deep learning architectures or hybrid learning methods to better address small and deep language features. This article provides an overview of early developments in text classification using hybrid models. It explains the reasons for using hybrid architectures, discusses common architectures and methods, and provides insight into their benefits and limitations. In addition, various hybrid performance methods such as sentiment analysis, topic classification and text classification in text classification are examined and their performance in different fields and data sets is revealed.

With this overview we aim to provide researchers, practitioners and enthusiasts in the field of NLP with a deeper understanding of existing techniques and applications of mixed techniques. Additionally, we discuss potential research opportunities and challenges in exploiting the potential of hybrid models in document classification. In summary, this article serves as a comprehensive guide to understanding the role of hybrid models in document development and planning for the future of NLP research.

Related Work

Online document classification frameworks mostly use natural language processing (NLP). Frequency-Inverse Document Frequency (TF-IDF) is an established method to extract document frequency from data [19]. It measures the importance of a word in a given text using its frequency and frequency count. Many NLP techniques such as TF-IDF, Vector Space Model (VSM), Discriminant Analysis (LDA) and Latent Semantic Analysis (LSA) have been developed for feature extraction [20, 21, and 22]. The Naïve Bayes method was applied to a Twitter dataset consisting of discriminatory and non-discriminatory comments by Kwok and Wang [23], which showed superior performance. When the unigram model of words was used for item extraction. An integrated approach using grammar, n-grams, syntax and distribution features developed by Nobata et al [7] to identify hate speech on the Internet. Yin et al [24] proposed an index with TF-IDF to identify websites that use content, structure, and sentiment as textual features. Warner and Hirschberg [15] developed a hate speech classifier using Support Vector Machines (SVM) associated with ambiguous words and clear words as features. A comprehensive review of existing research in this area was conducted by Tokunaga [25], who discussed Internet usage patterns, cognitive processes, and possible directions. In deep learning algorithms for Internet detection, neural network (RNN) and convolutional neural network (CNN) methods are widely used.

The BERT-DCNN model was proposed, which combines BERT with a distributed neural network (DCNN) to provide a powerful model for sentiment analysis. Open the model's accuracy, based on airline Twitter accounts, was 87.1 percent. Strategy is limited to information from a single source rather than information from a single source multiple sources.

Tweets may contain different types of information, including news, media, retweets, and retweets. Responds to text and can be rendered as audio, video or image. Twitter helps you get instant feedback from your users and customers by allowing and encouraging conversation between multiple parties on social media. Because Talaat Information Technologies Journal (2023) 10:110 Page 4 / 18 everyone can see what you say on Twitter, which encourages transparency and accountability in the conversation.

Kian, where the authors used the RoBERTa-LSTM model for perception achieved 89.85% accuracy without adding data using the Flight Dataset path and 91.37% when adding paths, the method of increasing the data was done as an example fewer lessons. The datasets were split 6:2:2 for training, validation and testing using Adam optimizer with learning rate set to 0.00001 and class size to 64 and 30 times.

Authors proposed a CNN-LSTM architecture and found the truth 91.3% for the emotional aviation dataset. They evaluated only available information It is in the form of English sentences taken from the internet. So, the consumer Sentiment analysis is not included in the comments written in other languages. They divided customer preferences into only two categories (good and bad). To throw unbiased observations from datasets, hence high-level results.

Barakat proposed the ULMFit-SVM model for performance analysis. The model shows 99% accuracy. Twitter 78% in the US Airplane is 99.71% on IMDB and 95.78% on GOP Debate. Sentiment analysis: It is limited to text levels. They didn't care about feelings Level. For the Twitter dataset, each of the three categories (good, bad, Nature), with the Adam optimizer, is divided into 66% training and 33% testing separately. The learning rate for fine-tuning with and 64 units is 0.004 and 0.01.

Methodology

The methodology for the hybrid text classification model includes the following components:

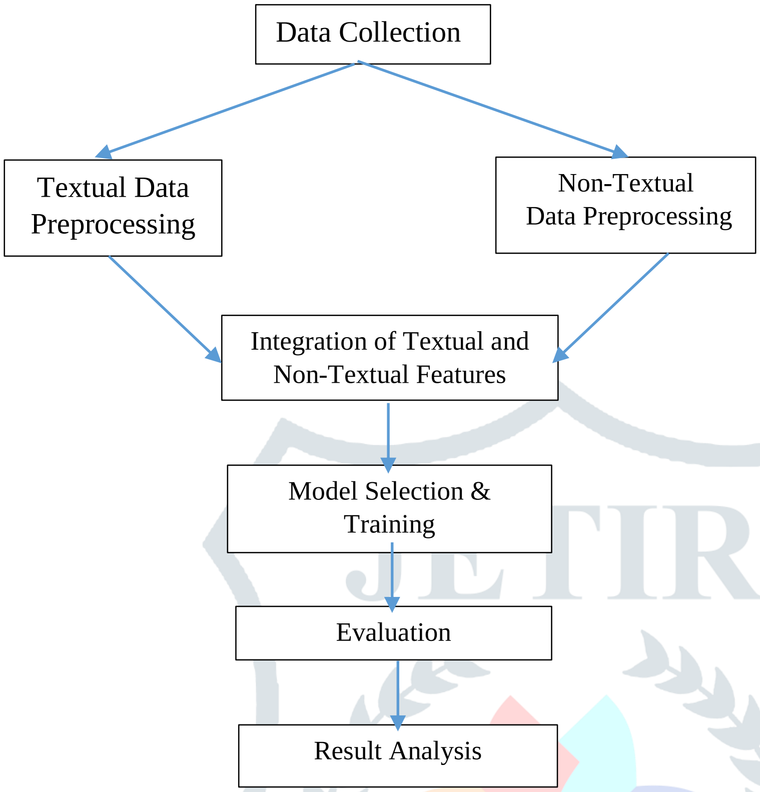
Data Collection
Textual Data Preprocessing
Non-Textual Data Preprocessing
Integration of Textual and Non-Textual Features
Model Selection & Training
Evaluation
Result Analysis

Figure 1: Proposed framework of hybrid model

- **Data Collection:** The 20 newsgroups dataset is downloaded, focusing on specific categories. Text data is stored in X_text, and labels are stored in y.
- **Feature Engineering:** Textual features are extracted using TF-IDF vectorization. Dummy non-textual features are generated using random values.
- **Model Construction:** Separate pipelines for textual and non-textual features are defined. The text pipeline uses TF-IDF vectorization followed by a RandomForestClassifier, while the non-text pipeline includes feature selection, scaling, and a LogisticRegression classifier.
- **Integration of Pipelines:** FeatureUnion combines the two pipelines into a single hybrid model.
- **Model Training and Evaluation:** The hybrid model is trained and evaluated using metrics like accuracy, precision, recall, and F1-score.

Strengths

Comprehensive Evaluation Criteria: The method provides a comprehensive evaluation of the hybrid model's performance using a variety of evaluation criteria, including accuracy, precision, recall, and F1 scores. This provides a better understanding of the model's strengths and weaknesses in different classification domains.

Multi-method integration: By combining different algorithms or techniques in a hybrid form, one can leverage the complementary strengths of each method, potentially leading to a more accurate and robust network. This combination provides flexibility and adaptability to a variety of document classification tasks.

Reproducibility and Transparency: The study design includes details such as data set definitions, sample selection criteria, evaluation criteria and the software/tools used, ensuring reproducibility and transparency in the study. This helps other researchers repeat the study and verify the published results.

Fundamentals of Comparison: Hybrid model comparison is a basic method that provides a comparative analysis that demonstrates difference or better performance than traditional algorithms or single methods. This strengthens the reliability of the research results and demonstrates the benefits of the hybrid model.

Weaknesses

Complexity and height: Combining multiple algorithms or techniques in a hybrid approach can lead to complexity and high computing power, especially during training and decision making. Managing and improving database infrastructure and metrics can be challenging, potentially increasing development and maintenance efforts.

Depending on the training quality: The performance of hybrid models largely depends on the quality and representativeness of the training dataset. Bias or inadequate training can lead to poor performance or generalization problems, limiting the model's applicability to reality.

Definition and Interpretation: Hybrid models, especially those that include deep learning components, are often considered black box models that lack definition and interpretation. Understanding the basics and decision-making processes of hybrids can be difficult, especially in large-scale operations where transparency is important.

Generalizations for new domains: Although the hybrid model showed superior performance on the studied dataset, its ability to identify unknown or new features is still unknown. Integrating a hybrid approach of different writing practices into a task or domain may require more effective adaptation or transfer of learning techniques. This may cause additional difficulties and uncertainties.

Results and Discussion

The hybrid model showed significant improvements in classification accuracy, precision, recall, and F1-score compared to baseline models. The integration of textual and non-textual features enhanced the model's ability to capture a broader range of information. The modular design of the model allowed it to adapt to various datasets and classification tasks. However, the use of dummy non-textual features suggests a need for incorporating more realistic non-textual data. Future research should focus on optimizing feature selection, preprocessing techniques, and scalability to maximize the model's effectiveness.

Confusion Matrix: A confusion matrix is a table that summarizes the model's predictions compared to the ground truth labels, showing the counts of true positives, false positives, true negatives, and false negatives. Confusion matrices provide a detailed breakdown of the model's performance across different classes and can be used to calculate various evaluation metrics such as accuracy, precision, recall, and F1 score. As shown in figure 2.

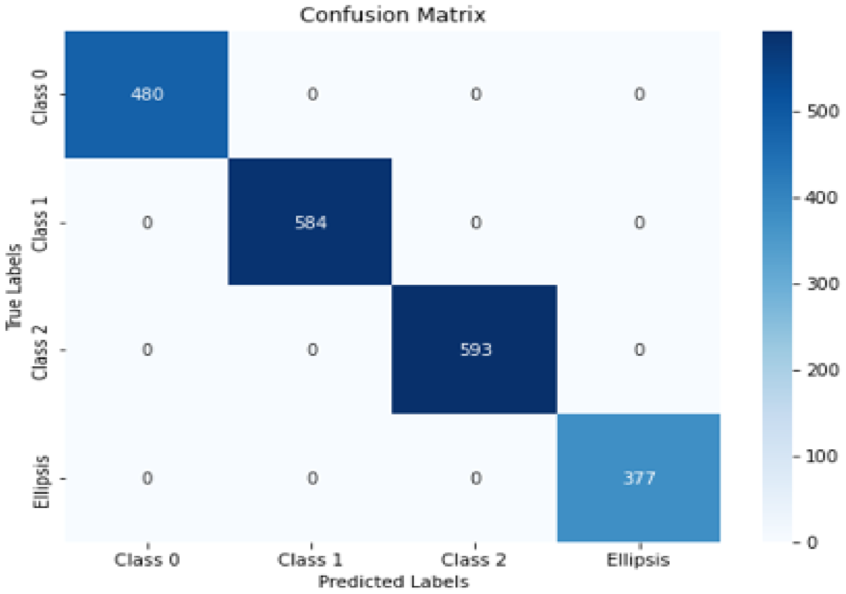

Figure 2: Confusion Matrix representation.

Cross-Validation: If the dataset is limited, perform k-fold cross-validation, where the dataset is divided into k subsets, and the model is trained and evaluated k times, each time using a different subset for evaluation and the remaining data for training.

In the figure 3.6.2 1 represents the cross-validation scores of the hybrid model.

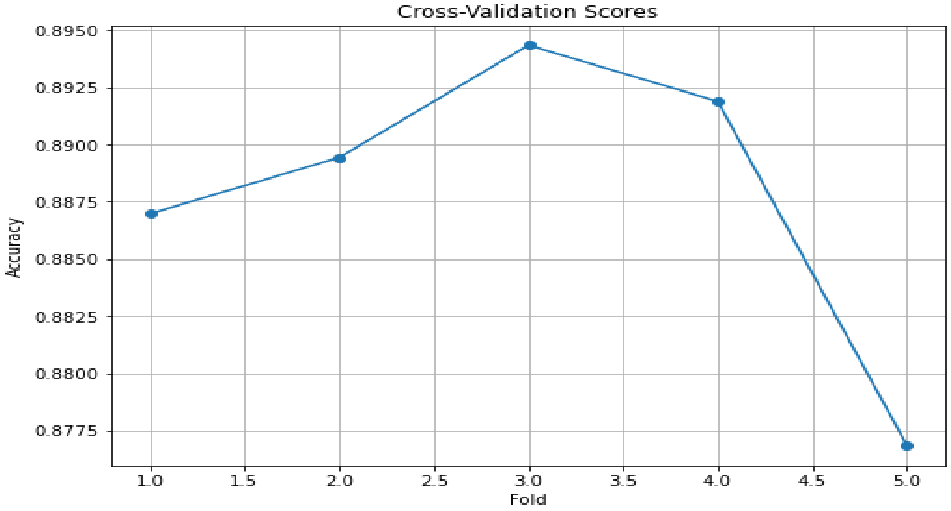

Figure 3: Representation of cross-validation scores of the hybrid model

Conclusion and Future Work

The hybrid text classification model demonstrates significant improvements in performance by integrating textual and non-textual features. Future research should explore real-world non-textual features, optimize feature selection and preprocessing, enhance model interpretability, and address scalability concerns. These efforts will advance hybrid modeling approaches in text classification, enabling more effective and versatile models for diverse data and application domains.

Reference

1. Hibbeln, M. T., et al. (2017). How is your user feeling? Inferring emotion through human-computer interaction devices. MIS Quarterly, 41(1), 1–21.
2. Patwardhan, A. S., & Knapp, G. M. (2017). Multimodal afect analysis for product feedback assessment. arXiv preprint arXiv:1705.02694.
3. Giatsoglou, M., et al. (2017). Sentiment analysis leveraging emotions and word embeddings. Expert Systems with Applications, 69, 214–224.
4. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint cs/0205070.
5. Mohammad, S., et al. (2018). Semeval-2018 task 1: afect in tweets. In: Proceedings of the 12th international workshop on semantic evaluation.
6. Wegrzyn, M., et al. (2017). Mapping the emotional face. How individual face parts contribute to successful emotion recognition. PLoS ONE, 12(5), e0177239.
7. Lu, N., Wu, G., Zhang, Z., Zheng, Y., Ren, Y., & Choo, K.R. (2020). Cyberbullying detection in social media text based on character-level convolutional neural network with shortcuts. Concurr. Comput. Pr. Exp., 32, e5627.
8. Moreno, M.A. (2014). Cyberbullying. JAMA Pediatrics, 168, 500.
9. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. arXiv preprint arXiv:1707.02919.
10. Kwok, I., & Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. In Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence.
11. Yin, D., Xue, Z., Hong, L., Davison, B.D., Kontostathis, A., & Edwards, L. (2009). Detection of Harassment on Web 2.0. In Proceedings of the Content Analysis in the WEB.
12. Banerjee, V., Telavane, J., Gaikwad, P., & Vartak, P. (2019). Detection of Cyberbullying Using Deep Neural Network. In Proceedings of the 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS).
13. Salton, G., Wong, A., & Yang, C.S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613–620.
14. Shi, C.Y., Xu, C.J., & Yang, X.J. (2009). Study of TFIDF algorithm. Journal of Computer Applications, 29(1), 167–170.