### Table of contents

1. Introduction Background

Introduction to text classification method

1. Related Work

Table 1. State-Of-The-Art Models for text classification Table 2. Different dataset in news classification domain

1. Dataset

Table 3. Ministries and their corresponding label that we have considered in the dataset

Table 4. Some of the sample from our dataset Statistics of our dataset

1. Model Architecture and hyper-parameters Our applied methods

Architecture Variants

1. Experiments and results Performance Measure
2. Conclusion
3. Future Scope

References

# Introduction

The proliferation of fake news, especially on social media platforms, has become an increasing concern in recent years when it comes to its effects on governmental

institutions and processes. Fake news often refers to intentionally false or misleading content presented as real news meant to influence public opinion or obscure the truth.

With the rise of social media and online news, fake news has enabled the rapid spread of misinformation and disinformation to large audiences. This flood of false information has made it difficult for citizens to differentiate truth from fiction when forming opinions about political leaders, parties, and policy issues. It has also

introduced more confusion and polarization into political discourse.

Experts warn that the unchecked spread of fake news could undermine the functioning of democracy by decreasing public trust in societal institutions. When citizens cannot even agree on basic facts due to the distortion of truth, it becomes challenging to have reasoned debates on policy decisions. There are also fears that the proliferation of fake news makes elections more vulnerable to external influence via targeted misinformation campaigns aimed at specific groups of voters.

In response, world governments have grappled with solutions, including considering regulations for social media platforms, public education programs, and third-party fact-checking initiatives. However, identifying an appropriate and ethical societal-level approach has remained a complex challenge given concerns over limiting freedom of speech. For now, the spread of misinformation through fake news continues to pose real threats to informed civic participation, election integrity, and governance credibility until more robust sociotechnical systems can be implemented.

<TODO - STATEMENT BY THE FACEBOOK REGARDING THE INTERFERENCE OF CHINA, AND THE SPENDING AND ACTIONS OF THE GOVERNMENT THAT ARE BEEN TAKEN>

Thus in this paper we present a state-of-the-art method for classifying the news as per the specific ministry to which it belongs in the government, And than giving that news to the specific ministry of the government to classify that news as the fake or not.

This study aims to take a different technical approach to the problem of the fake news and presents a state-of-the-art model for news classification based on the different ministries in that specific country and than gives that specific news to the ministries to whom regarding that news is to classify, And if that specific news is fake than that ministry of the government can take further actions regarding that news to eradicate the problem of mis-information from the country.

## Background

The murky web of fake news entangling governments in a tangled mess has become a significant concern in recent years. The pervasive threat of fake news has woven a complex and intricate web, casting a long shadow on the legitimacy and effectiveness of governments worldwide. This web of misinformation acts as a corrosive agent, eroding public trust in elected officials and government institutions. They cause the following damage to the governments:

**Erode Public Trust:** Fake news acts as a corrosive agent, eating away at the public's confidence in its elected officials. Fabricated stories about corruption, scandals, or policy failures, if widely spread, can sow seeds of doubt and suspicion. Imagine headlines screaming about government officials secretly plotting against their own citizens, even if demonstrably false, can damage public perception and make genuine government initiatives seem tainted.

**Hinder Decision-Making:** When the information landscape is polluted with misinformation, making informed decisions becomes a herculean task. Policymakers navigating complex issues like healthcare or economic reform rely on accurate data and public feedback. Fake news can skew both, leading to policies based on false premises and public backlash towards decisions with legitimate reasoning, but obscured by a fog of fabricated narratives.

**Polarize and Incite:** Perhaps the most dangerous effect of fake news is its ability to exacerbate existing divisions and incite animosity. Slanted stories targeting specific groups can fuel social and political polarization, creating echo chambers where people are exposed only to information that confirms their pre-existing biases. This can lead to real-world consequences, from protests fueled by falsehoods to acts of violence driven by hate mongering disguised as news.

However, this is a complex battle with no easy solutions. Balancing freedom of speech with the need to curb harmful misinformation is a delicate dance. Finding effective solutions requires collaboration between governments, technology companies, media organizations, and citizens themselves.

Ultimately, the fight against fake news is a fight for the soul of democracy. Just as the printing press revolutionized how information was shared, the digital age has democratized it, presenting both immense opportunities and significant challenges.

Advances in machine learning and deep learning have opened up new possibilities for automatically categorizing news content by the government ministry or ministries

it is most relevant to. By feeding these AI systems large datasets of news articles that have already been manually assigned ministry tags, they can learn to make these categorizations on their own with a high degree of accuracy.

For example, a news article discussing changes in educational policy would be tagged as belonging to the Ministry of Education. One about infrastructure spending may be designated for both the transportation and construction ministries. Over time, deep learning algorithms can discern text patterns that correlate with specific ministries based on word choice, names mentioned, legislation references, and more.

Powerful neural network architectures could process an incoming news article and almost instantly predict the most probable ministry or ministries it should be routing to. This could occur automatically behind the scenes, ensuring relevant officials and analysts are kept informed of developments that concern their areas of governance.

As more labeled data is provided, the machine learning models can grow even more sophisticated at multi-label classification. They may develop subtler text understandings and model interdependencies between ministries. These AI systems could become versatile automated assistants that relieve some manual categorization burdens in government newsrooms and communication teams. Their

integration promises to bring news filtering and dissemination firmly into the era of AI- augmented governance.

The rapid advances in this type of content-based predictions promise more targeted, responsive flows of information across all levels of government. This has far-reaching implications for connecting policies to accurate timely data.

## Introduction to text classification method

## Text classification is a technique used in natural language processing (NLP) to automatically categorize text documents into predefined categories or classes. It involves training a machine learning model or neural network to learn patterns and features from a labeled dataset and then using the trained model to predict the category of new, unseen text data.

Traditionally, text classification models relied on machine learning algorithms such as Support Vector Machines (SVM), Naive Bayes, or Decision Trees. These models required handcrafted features to represent the text data, such as bag-of-words or TF- IDF vectors. While these approaches achieved reasonable accuracy, they often struggled with capturing complex linguistic patterns and semantic relationships in the text.

The advent of deep learning revolutionized text classification by enabling models to automatically learn hierarchical representations of text data. Instead of relying on handcrafted features, deep learning models, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), or Transformers, can learn feature representations directly from raw text inputs.

Deep learning models for text classification have shown remarkable performance improvements over traditional machine learning approaches. They can capture

intricate linguistic patterns, semantic relationships, and contextual information, resulting in more accurate and robust text classification.

Multi-label text classification extends the traditional text classification task by allowing documents to belong to multiple categories simultaneously. Instead of assigning a single label to a document, the model predicts a set of labels, each representing a different category. This is particularly useful when dealing with text data that can belong to multiple topics or have overlapping themes.

Deep learning-based models, such as multi-label CNNs or multi-label Transformers, have been successfully applied to multi-label text classification tasks. These models leverage their ability to learn complex representations and capture dependencies between different labels, resulting in accurate predictions for multiple categories.

In recent years, the field of text classification has seen advancements driven by deep learning techniques. Researchers have explored novel architectures, such as hierarchical models, attention mechanisms, and pre-trained language models like

BERT, GPT, and RoBERTa. These models have achieved state-of-the-art results on various text classification benchmarks and have become the current trend in research and industry applications.

Overall, the revolution from traditional machine learning to deep learning has significantly improved text classification performance and opened up new possibilities for solving complex NLP tasks. By leveraging the power of deep learning models, we can achieve more accurate and robust text classification results, enabling applications in various domains such as news classification, sentiment analysis, and document organization.

# Related Work

In order to create the comprehensive table presented in this document, an extensive literature review was conducted. This involved analyzing various methods

implemented by other researchers in the field of text classification. The table includes the title, authors, year, and a brief description of each method, providing a valuable resource for identifying the most accurate and suitable method for the research task at hand. The literature review process involved carefully selecting relevant papers and extracting key information from them. This enabled the authors to gather a diverse range of approaches and insights, ensuring a well-rounded and informed analysis. The resulting table serves as a reference point for researchers and practitioners interested in the latest advancements in text classification.

## Table 1. State-Of-The-Art Models for text classification

|  |  |  |  |
| --- | --- | --- | --- |
| Title | Authors | Year | Brief Description |
| Political News Bias Detection using Machine Learning [1] | Minh Viu | 2017 | In this paper they have used the MLP (multi layer perceptron) approach to  identify the sentimental bias of the given text. They used the MLP layer with different set of parameters. |
| BerConvoNet: A deep learning framework for fake news classification [2] | Choudhary et. al. | 2021 | They presented the  BerConvoNet a deep learning framework to classify the given news text into real or fake, The BerConvNet has two main building blocks: a news embedding block (NEB) and a |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | multi-scale feature block (MSFB). |
| Research on hot news classification algorithm based  on deep learning [3] | Wang et. al. | 2019 | This paper studies deep  learning application in text categorization, combined with the characteristics of the news text, and put forward the double Bi-Gated Recurrent  Unit (GRU) + attention deep learning model to predict hotspots, and achieved good results. |
| Hybrid Approach  Combining Machine Learning and  a Rule-Based Expert System for Text  Categorization [4] | Villena-Román et. al. | 2019 | A novel hybrid approach for text categorization that combines a machine learning algorithm, which provides a base model trained with a  labeled corpus, with a rule- based expert system, which is used to improve the results provided by the previous classifier by filtering false positives and dealing with false negatives is discussed. |
| Different Machine Learning based Approaches of  Baseline and Deep Learning Models for Bengali News  Categorization [5] | Rabib Hossain et. al. | 2020 | The paper discusses the application of different machine learning approaches, including baseline models like naive-bayes and logistic regression, as well as deep  learning models like CNN and Bi-LSTM, for Bengali news categorization. |
| Fake News  Classification Based on Content Level Features [6] | Lai et. al. | 2022 | The combination of ML and NLP are implemented to classify fake news based on an open, large and labeled corpus on Twitter and it is found that the neural network models outperform the traditional ML models by, on average, approximately 6% |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | precision, with all Neural network models reaching up to 90% accuracy. |
| Automated Amharic News Categorization Using Deep Learning Models [7] | Endalie et. al. | 2021 | In this article, the authors used fastText to generate text vectors to represent semantic meaning of texts and solve the problem of traditional methods, and the text vectors matrix is then fed into the embedding layer of a convolutional neural network (CNN), which automatically extracts features, And model achieves the classification accuracy of the 93.79%. |
| Deep Learning–based Text Classification: A Comprehensive  Review [8] | Minaee et. al. | 2021 | A comprehensive review of more than 150 deep learning based models for text classification developed in recent years is provided, and their technical contributions, similarities, and strengths are discussed. |
| Deep Pyramid  Convolutional Neural Networks for Text  Categorization [9] | Johnson et. al. | 2017 | The paper proposes a low- complexity deep pyramid CNN model for text categorization that efficiently represents long- range associations in text and outperforms previous models on benchmark datasets. |
| FASTTEXT.ZIP : COMPRESSING TEXT  CLASSIFICATION MODELS [10] | Joulin et. al. | 2016 | This work proposes a method built upon product quantization to store the word embeddings, which produces a text classifier, derived from the fastText approach, which at test time requires only a fraction of the memory compared to the original one, without noticeably sacrificing |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | the quality in terms of classification accuracy. |
|  |  |  | This article constructed |
| Character-level Convolutional  Networks for Text Classification∗ [11] | Zhang et. al. | 2015 | several large-scale datasets to show that character-level convolutional networks could achieve state-of-the-art or  competitive results in text |
|  |  |  | classification. |

An extensive literature review was conducted to identify various news classification datasets for this study. The review aimed to gather comprehensive information about the different dataset used by the researches for news classification. The resulting table includes the dataset name and brief description of each dataset, providing a valuable resource for identifying the most relatable and suitable dataset for the research work.

## Table 2. Different dataset in news classification domain

|  |  |
| --- | --- |
| Dataset | Brief Description |
| Reuters News [14] | The Sogou News dataset is a mixture of the SogouCA and SogouCS news corpora. The classification labels of the news are determined by their domain names in the URL.  For example, the news with URL  [http://sports.sohu.com](http://sports.sohu.com/) is categorized as a sport class. |
| Sogou News [13] | The Sogou News dataset is a mixture of the SogouCA and SogouCS news corpora. The classification labels of the news are determined by their domain names in the URL.  For example, the news with URL  [http://sports.sohu.com](http://sports.sohu.com/) is categorized as a sport class. |
| AG News [12] | The AG News dataset is a collection of news articles collected from more than 2,000  news sources by ComeToMyHead, an academic news search engine. This dataset includes 120,000  training samples and 7,600 test samples. Each sample is a short text with a  four-class label. |
| Newsgroups [15] | The 20 Newsgroups dataset is a collection of newsgroup documents posted on 20 different topics. Various versions of this dataset are used for text classification, text  clustering and so one. One of the most popular versions contains 18,821 |

|  |  |
| --- | --- |
|  | documents that are evenly classified across all topics. |

Other datasets developed for news categorization includes: BBC [18], Google news [17], Bing news [16].

# Dataset

### (if not works try 4th one of whole zero class used in sample model)

As part of our objective to classify news according to the ministries of the Indian government, we encountered a lack of available datasets. To address this, we took the initiative to create our own dataset for this research endeavor.

The dataset was constructed by merging three news datasets from Kaggle and subsequently manually labeling the news according to the corresponding ministries. It is important to note that while there are over 100 ministries in the Indian government, we were able to consider 59 ministries due to limitations in data availability and other constraints. and the 60th number/label was given to the no ministry, as the cases are possible where any news does not belongs to any of the ministry of Indian goverenment.

We have made this dataset openly accessible under the MIT license, and it can be downloaded from the following Kaggle link: <https://www.kaggle.com/neelshah2022/code>. We kindly request that this dataset be utilized ethically and for the betterment of society.

The dataset comprises of 36051 news, And 60 classification classes or labels. With no duplicate news in the dataset.

Here is the list of ministries and their corresponding label numbers that were used in the creation of the dataset:

**Table 3. Ministries and their corresponding label that we have considered in the dataset**

|  |  |
| --- | --- |
| Label | Ministry |
| 0 | Ministry of Railways |
| 1 | Ministry of Rural Development |
| 2 | Ministry of Steel |
| 3 | Ministry of Science & Technology |

|  |  |
| --- | --- |
| 4 | Ministry of Information & Broadcasting |
| 5 | Ministry of Food Processing Industries |
| 6 | Ministry of Health and Family Welfare |
| 7 | Ministry of Human Resource Development |
| 8 | Ministry of Agriculture |
| 9 | Ministry of Environment and Forests |
| 10 | Ministry of Chemicals and Fertilizers |
| 11 | Ministry of Water Resources |
| 12 | Ministry of Defence |
| 13 | Ministry of Petroleum & Natural Gas |
| 14 | President's Secretariat |
| 15 | Ministry of Micro, Small & Medium Enterprises |
| 16 | Ministry of Mines |
| 17 | Ministry of Tourism |
| 18 | Ministry of Housing & Urban Affairs |
| 19 | Ministry of Coal |
| 20 | Prime Minister's Office |
| 21 | Ministry of Textiles |
| 22 | Ministry of Commerce & Industry |
| 23 | Ministry of External Affairs |
| 24 | Ministry of Social Justice & Empowerment |
| 25 | Ministry of Power |
| 26 | Ministry of Consumer Affairs, Food & Public Distribution |
| 27 | Ministry of Heavy Industries & Public Enterprises |
| 28 | Ministry of Communications |
| 29 | Ministry of Shipping |
| 30 | Ministry of Finance |
| 31 | Ministry of Tribal Affairs |
| 32 | Ministry of Statistics & Programme Implementation |
| 33 | Ministry of Labour & Employment |
| 34 | Ministry of Law & Justice |
| 35 | Vice President's Secretariat |

|  |  |
| --- | --- |
| 36 | Ministry of Civil Aviation |
| 37 | Ministry for Development of North-East Region |
| 38 | UPSC |
| 39 | Ministry of Agro & Rural Industries |
| 40 | Ministry of Home Affairs |
| 41 | Ministry of Youth Affairs and Sports |
| 42 | Special Service and Features |
| 43 | Ministry of New and Renewable Energy |
| 44 | Ministry of Parliamentary Affairs |
| 45 | Planning Commission |
| 46 | Ministry of Personnel, Public Grievances & Pensions |
| 47 | Election Commission |
| 48 | Department of Space |
| 49 | Ministry of Disinvestment |
| 50 | Department of Ocean Development |
| 51 | Ministry of Overseas Indian Affairs |
| 52 | Ministry of Housing and Urban Poverty Alleviation |
| 53 | Ministry of Culture |
| 54 | Ministry of Company Affairs |
| 55 | Ministry of Panchayati Raj |
| 56 | Cabinet Committee on Economic Affairs (CCEA) |
| 57 | Cabinet |
| 58 | Department of Atomic Energy |
| 59 | Cabinet Committee Decisions |
| 60 | No Ministry |

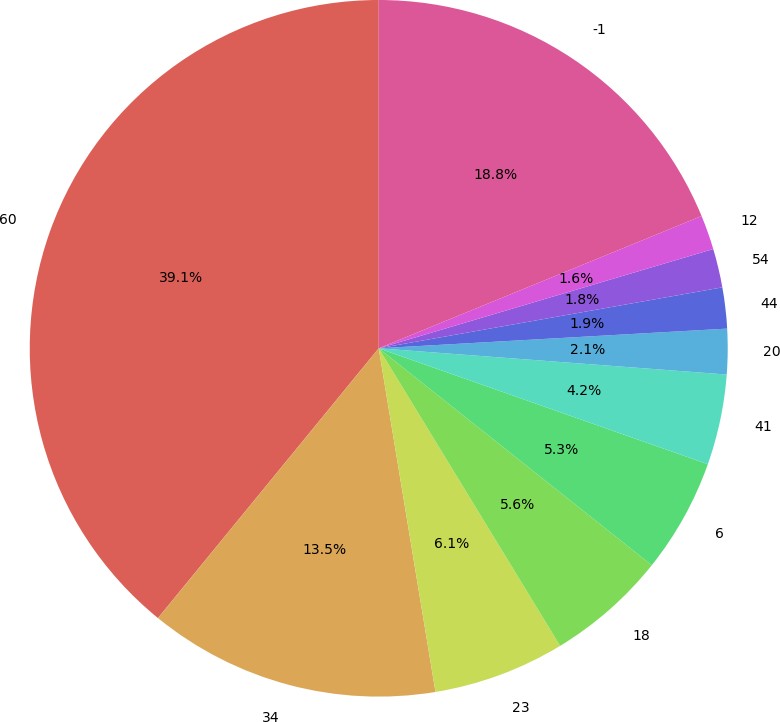
Below is a table showcasing some examples from the dataset for demonstration purposes. These examples include the news title, news content, and the corresponding ministries associated with each news article.

**Table 4. Some of the sample from our dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| News Id | News Title | News Content | Ministries Associated |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | A 35-year-old State Reserve |  |
|  |  | Police Force (SRPF) jawan, |  |
|  |  | identified as Shrikant Berad, |  |
|  |  | allegedly shot his 33-year-old |  |
| 24216 | SRPF jawan shoots colleague dead, kills self in Maharashtra's Gadchiroli | colleague dead before killing himself with his service rifle in Maharashtra's Gadchiroli district, police said. The  incident took place on  Wednesday at Marpalli police | Ministry of Defence (12) |
|  |  | post in Aheri tehsil, police |  |
|  |  | added. The motive behind the |  |
|  |  | incident is being ascertained, |  |
|  |  | police further said. |  |
|  |  | England all-rounder Moeen Ali, |  |
|  |  | who announced retirement |  |
|  |  | from Test cricket last year, has |  |
|  |  | revealed that England's newly- |  |
|  | McCullum | appointed Test coach Brendon |  |
|  | messaged me | McCullum has asked him if he | Ministry of Youth |
| 24226 | asking if I was in: | would be available to play | Affairs and Sports |
|  | Moeen on Test | Tests again in future. "We | (41) |
|  | cricket comeback | spoke...[he asked] would I be |  |
|  |  | available [in the future]? I said, |  |
|  |  | 'Call me at the time'. We'll |  |
|  |  | see...The door is open," stated |  |
|  |  | Moeen. |  |
|  |  | At least two labourers died | Ministry of Health  (7) and Family Welfare, Ministry of Labour &  Employment (33) |
|  |  | after soil collapsed on them at |
|  |  | an under-construction plot in |
|  |  | Gurugram's Sector 57 on |
|  |  | Wednesday, police said. "The |
|  | 2 labourers die as | basement had been dug over |
| 24263 | soil collapses at  construction site in | 10 feet below the ground  level...labourers were putting |
|  | Gurugram | pillars...when a mound of soil |
|  |  | collapsed on them...they were |
|  |  | trapped underneath," police |
|  |  | added. It took over 15 minutes |
|  |  | to retrieve their bodies, police |
|  |  | stated. |

## Statistics of our dataset



Where the -1 indicates the remaining ministries. Other than that shown here, And the numbers outside the pie are the ministry index number by which that specific ministry was classified in the dataset.

# Model Architecture and hyper- parameters

## Our applied methods

Our neural network model is designed for natural language processing (NLP) text classification. It consists of an embedding layer for word representation, two bidirectional LSTM layers for capturing contextual information, and two dense layers for classification. The model is compiled for multi-class classification using categorical crossentropy as the loss function and the Adam optimizer with a learning rate of 0.0010. The MirroredStrategy is employed for parallel training on multiple GPUs. We can use the Top 3 or Top 5 accuracy to make predictions about the which class should the model predict.

**Architecture Variants**

# Experiments and results

## Performance Measure

Here are some of the key performance measures used for multi-class text classification problems in natural language processing:

Accuracy: Fraction of predictions the model got right out of all predictions made. Provides an overall measure of how many times the classifier was correct.



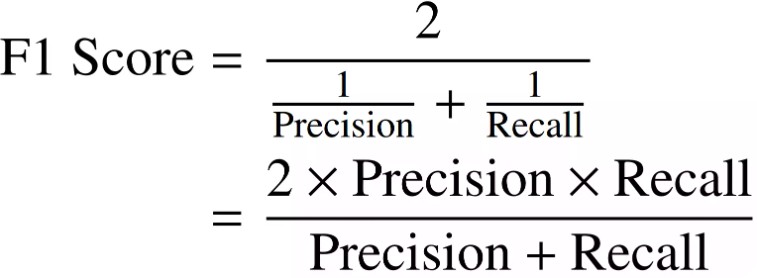
Precision: For a specific class, precision measures the proportion of predicted positive examples that actually belong to that class. It provides an understanding of false positives.



Recall: This measures the proportion of correctly detected examples out of all the examples in a specific class. It is useful for identifying instances where the model incorrectly fails to detect something.



F1-score: Harmonic mean of precision and recall. It is particularly useful when classes are imbalanced.



Top 3

Top 5

Top 10

**Results**

# Conclusion

Our model provides ministries with personalized news content tailored to their specific areas of interest. Ministries have the ability to distinguish articles as either 'fake' or 'not fake' in order to effectively combat the spread of misinformation. Additionally, our model offers data-driven insights to support

informed policy-making and decision-making processes.

Furthermore, our model fosters collaboration and facilitates discussion among different ministries, enabling them to work together more effectively. This promotes a more cohesive and coordinated approach to governance.

In addition to combating false information, our model also plays a crucial role in emergency situations. It allows for quick identification and distribution of critical news during times of crisis, ensuring that accurate and timely information reaches the public.

Finally, our model helps shape communication strategies by providing valuable

insights into public sentiment and media coverage. This allows ministries to tailor their messaging and engagement strategies to effectively reach their target audience and achieve their communication objectives.

# Future Scope

In order to enhance the precision and scale of this model, it is recommended to utilize a larger dataset. Currently, the model has been developed using a small

dataset of approximately 36,000 news articles.

It is important to note that this model has only been designed for the 59 ministries of India, despite the existence of more ministries in the country. Therefore, it is suggested to expand the model by incorporating a dataset that includes a wider range of ministries.

While this model has been tailored specifically for the Indian government, the concept can be extended to other democratic countries such as the United

States of America, United Kingdom, Germany, and many others.

Another potential area for future exploration is to analyze the impact of

implementing this model in the country and evaluate the advantages and disadvantages it brings to society.

# References

|  |  |
| --- | --- |
| Num | Citation |
| 1 | Vu, M. (2017). Political news bias detection using machine learning. URL: [https://pdfs](https://pdfs/). semanticscholar. org/8445/2eb068bdfe7d5809734a5da8f5c7d10bebfa. pdf. |
| 2 | Choudhary, M., Chouhan, S. S., Pilli, E. S., & Vipparthi, S. K. (2021). BerConvoNet: A deep learning framework for fake news classification. Applied Soft Computing, 110, 107614. |
| 3 | Research on hot news classification algorithm based on deep learning Zaiying Wang, Bohao Song Xi’an University of Science & Technology, Xi’an, Shaanxi, 710054, China |
| 4 | Wang, Z., & Song, B. (2019, March). Research on hot news classification algorithm based on deep learning. In 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) (pp. 2376-2380). IEEE. |
| 5 | Hossain, M. R., Sarkar, S., & Rahman, M. (2020). Different machine learning based approaches of baseline and deep learning models for bengali news categorization. International Journal of Computer  Applications, 975, 8887. |
| 6 | Lai, C. M., Chen, M. H., Kristiani, E., Verma, V. K., & Yang, C. T. (2022). Fake news classification based on content level features. Applied  Sciences, 12(3), 1116. |
| 7 | Endalie, D., & Haile, G. (2021). Automated Amharic news categorization using deep learning models. Computational Intelligence and Neuroscience, |

|  |  |
| --- | --- |
| Num | Citation |
|  | 2021. |
| 8 | Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning--based text classification: a comprehensive review. ACM computing surveys (CSUR), 54(3), 1-40. |
| 9 | Johnson, R., & Zhang, T. (2017, July). Deep pyramid convolutional neural networks for text categorization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 562-570). |
| 10 | Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. arXiv preprint arXiv:1612.03651. |
| 11 | Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. Advances in neural information processing systems, 28. |
| 12 | Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. Advances in neural information processing systems, 28. |
| 13 | Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification?. In Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019,  Proceedings 18 (pp. 194-206). Springer International Publishing. |
| 14 | Reuters. [n.d.]. Retrieved from <https://martin-thoma.com/nlp-reuters>. |
| 15 | 20 Newsgroups. [n.d.]. Retrieved from <http://qwone.com/jason/20Newsgroups/>. |
| 16 | Wang, F., Wang, Z., Li, Z., & Wen, J. R. (2014, November). Concept-based short text classification and ranking. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (pp. 1069-1078). |
| 17 | Greene, D., & Cunningham, P. (2006, June). Practical solutions to the problem of diagonal dominance in kernel document clustering. In  Proceedings of the 23rd international conference on Machine learning (pp.  377-384). |
| 18 | Das, A. S., Datar, M., Garg, A., & Rajaram, S. (2007, May). Google news personalization: scalable online collaborative filtering. In Proceedings of the 16th international conference on World Wide Web (pp. 271-280). |
| 19 | CNN-News18 Headlines with Descriptions. [n.d.]. Retrieved from [https://www.kaggle.com/datasets/mauryansshivam/cnn-news18-headlines-](https://www.kaggle.com/datasets/mauryansshivam/cnn-news18-headlines-with-descriptions) [with-descriptions](https://www.kaggle.com/datasets/mauryansshivam/cnn-news18-headlines-with-descriptions). |

|  |  |
| --- | --- |
| Num | Citation |
| 20 | News Articles Dataset from Indian Express. [n.d.]. Retrieved from [https://www.kaggle.com/datasets/pulkitkomal/news-article-data-set-from-](https://www.kaggle.com/datasets/pulkitkomal/news-article-data-set-from-indian-express) [indian-express](https://www.kaggle.com/datasets/pulkitkomal/news-article-data-set-from-indian-express). |
| 21 | NEWS SUMMARY. [n.d.]. Retrieved from <https://www.kaggle.com/datasets/sujalbhagathansda/news-summary/>. |
|  |  |