BBC News Classification using CNN-GloVe & LSTM-Word2Vec models

Nidhi Joshi
230017936
Data Science (MSc) – FT
nidhi.joshi@city.ac.uk
(Colab file link: https://colab.research.google.com/drive/1imns3QRVqXaIn2rHw07qDilns00fJxmC?usp=sharing)

1 Problem statement and Motivation

The exponential rise in digital news consumption has led to a vast information accessible online, posing a significant challenge for the end consumers seeking relevant news content. The sheer volume of daily articles published by news organizations, including the BBC, has made it increasingly difficult for individuals to navigate and access the news that aligns with their interests. With this overload of information, it becomes more important for publications to personalize the content for each individual and allow for efficient navigations and content filtering across their websites.[1]

In this project, we aim to address this challenge by developing a news classification system that categorizes BBC news articles into specific topics using and comparing performances of GloVe and Word2Vec-LSTM models for the purpose.

The motivation behind the endeavour is to provide the end-user with an enhanced experience of getting customized news articles filtered as per their interest. Filtering would be time-saving and enhance efficiency resulting in better user engagement. Not only that, but even the publications can reap the benefits through this system by analysing and getting insights on different consumer behaviours across different geographies. We can define our research question as - To what extent can advanced natural language processing techniques, such as GloVe and Word2Vec-LSTM models improve the efficiency and accuracy of BBC News classification thereby, enabling personalized news consumption in the era of information overload?

2 Research hypothesis

The research hypothesis is that for this particular task of categorising BBC news articles, GloVe will result in an efficient model owing to its ability to capture the global co-occurrence statistics of words within a corpus through word vectors.[2] We hypothesize that while both GloVe and Word2Vec-LSTM are known to capture semantic relationships between words and preserve semantic similarities through dense vectors, GloVe is better in word-level representations and hence would produce better results for this task. GloVe word embeddings would facilitate the article categorization into business, entertainment, politics, sport, and technology by understanding the meanings of the words. We would employ both the techniques for our analysis to test this hypothesis and arrive at the best solution for our posed research question.

3 Related work and background

News classification and categorization have been extensively studied in the field of natural language processing (NLP) and machine learning. In this section, we provide an overview of relevant research in this area, drawing upon some past research papers.

Several studies have focused on the classification of BBC news articles into different categories. In the paper titled "News Classification and Categorization with Smart Function Sentiment Analysis"[3], the authors propose a smart search function embedded in a search engine using natural language toolkit (NLTK) and BM25 for indexing and pre-processing patterns. Their approach employs sentiment analysis techniques on BBC news data, achieving an accuracy rate of 85%. In

"Analysis of BBC News by Applying Classification Algorithms"[4], the authors compare three classification algorithms, including Multinomial Naive Bayes, K Nearest Neighbor's, and Stochastic Gradient Descent, for classifying news articles into categories. The results demonstrate high precision, with Stochastic Gradient Descent outperforming the other algorithms.

Deep learning models have also been explored for news classification tasks. In the study titled "A Robust Hybrid Approach for Textual Document Classification"[5], the authors propose a hybrid methodology that combines traditional feature engineering and deep learning techniques. They employ filter-based feature selection and a deep convolutional neural network, achieving significant improvements in classification accuracy on the 20 Newsgroups dataset and BBC news data.

The paper "News Articles Classification Using Random Forests and Weighted Multimodal Features"[6] addresses the same problem using a combination of textual and visual features. They extract N-gram textual features and generate visual features from representative images to improve classification accuracy. This work highlights the importance of incorporating multimodal features in news article classification tasks.

Another study "BBC news data classification using naïve bayes based on bag of words"[7], explores Naïve Bayes (NB) for text classification of BBC news data. They categorize news articles into different categories by employing various NB classifiers combined with feature extraction methods like TF-IDF. They achieved an impressive 97.6% accuracy through Complement classifier.

Furthermore, "Performing Data Augmentation Experiment to Enhance Model Accuracy: A Case Study of BBC News' Data"[8] addresses the challenge of text classification using traditional models and proposes data augmentation techniques. They utilize WordNet, a semantic lexical database, for augmenting textual data and enhance news data classification using an LSTM model. Experimental evaluations demonstrate promising results, with the LSTM model achieving higher accuracy from the augmented dataset as compared to the original one. The study "Text Classification of BBC News Articles and Text Summarization Using TextRank"[9] focuses on automatic text classification and text summarization using machine learning algorithms. The authors emphasize the importance of transforming natural language documents into numerical representations for training models. The study showcases the application of machine learning algorithms for text classification and summarization tasks. We have used a similar concept in our case by encoding values and using numerical representations of the data.

These research papers have covered various aspects around classification algorithms, deep learning models, multimodal features, data augmentation, sentiment analysis, and text summarization which were helpful in implementing some techniques in our models as well.

4 Accomplishments

- Task 1: Perform initial EDA on the dataset to analyse and transform the data if needed – Completed
- Task 2: Remove outliers, special characters and extra spaces, normalize unicode characters in the data– Completed
- Task 3: Parse the dataset and encode category labels - Completed
- Task 4: Perform decontraction, lemmatization to text – Completed
- Task 5: Perform EDA on the preprocessed dataset to analyse – Completed
- Task 6: Integer encode the input data to perform feature engineering (Tokenization) – Completed
- Task 7: Encode data into sequences and apply padding – Completed
- Task 8: Load GloVe vectors and identify missing words from our docs in the vectors - Completed
- Task 9: Split the pre-processed dataset into train, validation and test datasets – Completed
- Task 10: Define the CNN model architecture, train and fit – Completed

- Task 11: Evaluate the results from performance metrics - Completed
- Task 12: Apply Part-of-speech tagging along with padded sequences and evaluate the results again – Completed
- Task 13: Transform the dataset using StandardScaler and apply class weights to change the range of loss and evaluate results – Completed but failed to obtain results; in fact, accuracy fell to 35%
- Task 14: Re-shape the data, use appropriate sequence length and apply custom callbacks for parameter tuning – Completed
- Task 15: Increase convolutional layers in the model and evaluate again - Completed
- Task 16: With prepared sequences, define embedding matrix for training LSTM model utilising GloVe vectorizers – Completed
- Task 17: Define the model, compile and train using the defined callback classes and evaluate results – Completed
- Task 18: Increase LSTM and dense layers to check difference in results – Completed
- Task 19: Print model summaries and plot loss and accuracy curves - Completed

5 Approach and Methodology

Availability of a standard and balanced dataset contributed to establish an efficient pipeline using research studies done earlier as well. These helped me to arrive at the decision of trying the task with CNN-GloVe and LSTM-Word2Vec models. Data was cleaned after basic analysis of the dataset. After ensuring balanced distribution of classes, it was then pre-processed using tokenization, lemmatization and decontraction. To use GloVe vectors, the pre-trained data was downloaded[10] and pre-trained Word2Vec embeddings were utilised from 'Gensim' library. CNN model was defined with GloVe word embeddings with a number of convolutional layers and maxpooling, and LSTM model was designed with Word2Vec word embeddings as input. Both CNN and LSTM models used crossentropy loss functions and Adam optimiser. CNN model was trained with 'sparse categorical crossentropy' loss function as the targets are not hot-encoded but a list of integers. Early stopping techniques were used to prevent overfitting. Furthermore, models were evaluated and analysed through accuracy, precision, recall and F1 scores and by plotting loss and accuracy plots. With defined set of epochs and batch sizes, the process was monitored over each epoch and callback classes were used for hyperparameter tuning. Different learning rates were employed to experiment and obtain better results.

One major limitation faced was hyperparameter sensitivity. It took a very long time to find the optimal combination. However, callback classes for some of the parameters like learning rate helped achieve the results faster. Main libraries used to implement both the models were Tensorflow/Keras (for building and training the models), NumPy (for numerical computations), scikit-learn (for data pre-processing, performance metrics and hyperparameter tuning), Gensim (for training and using Word2Vec embeddings) and spaCy (for data pre-processing).

The major issue faced was that despite so many parameter tunings, the accuracy didn't improve for the LSTM model. The irony is that the results kept predicting for only one class until this report was written and the error realised – we were erroneously implementing binary classification in this case however, there were 5 classes. CNN-GloVe was correctly implemented and hence better results but due to lack of time, LSTM model could not be improved.

6 Dataset

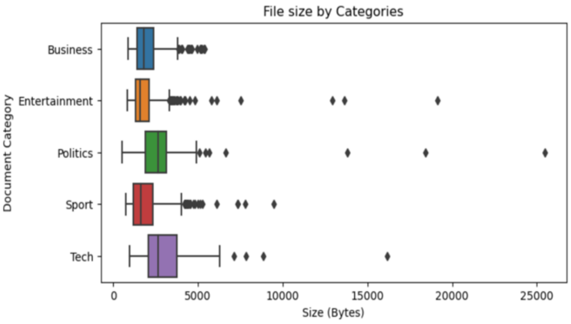
The BBC News Classification dataset available on Kaggle[11][12] is a collection of news articles from the BBC website. This data was originally sourced from the BBC news website corresponding to stories in five topical areas from 2004-2005[13]. It is widely used for text classification tasks, where the objective is to categorize the articles into predefined classes, vis. business, entertainment, politics, sport, and tech. The dataset includes news articles from various categories like:

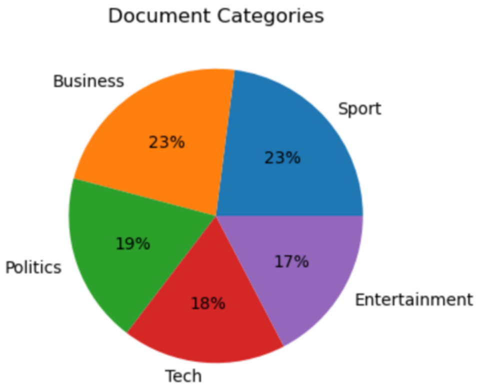
Category	Text

Business	UK economy facing major risks…
Entertainment	Musicians to tackle US red tape…
Politics	Baron Kinnock makes Lords debut...
Sport	Fuming Robinson blasts officials…
Tech	Mobiles rack up 20 years of use…

Table 1: Sample Category and text

We observed from initial EDA of the dataset that there were 2225 total records classified into 5 main categories divided in the following way: Business – 381, Entertainment – 510, Politics – 506, Sport – 413 and Tech – 395. Although the dataset was balanced, news articles can contain ambiguous language or can mention multiple topics within one article making it difficult to assign to a single class. For the same purpose, cosine similarity was implemented in the case of GloVe vectors which assigns the similarity between two words ranging from 0 to 1 – 0 indicating minimum distance (high similarity) and closer to 1 indicates maximum distance (low similarity).





Although this dataset deals with broader categories which are distinguishable from each other but in case of closely related categories like 'business' and 'finance' or 'entertainment' and 'culture', the classification would become a difficult task. The GloVe vectors were sourced from Kaggle[14]. A larger vector file dataset could've potentially enhanced the model performance, but for the purpose of this project, we employed this size for now.

1) Dataset Pre-processing

Obtained data was cleaned before the pre-processing. The outliers, special characters and extra spaces were removed to avoid unnecessary noise. The variations in character representation were handled by normalizing the Unicode characters and in vocabulary were handled by Decontraction. Lemmatization was employed to reduce words to their base form, removing whitespace, lowercasing the words and checking on stop words and punctuation. This helped in reducing vocabulary size and handling different word forms. Furthermore, Part-of-Speech tagging was used to assign grammatical tags to words enhancing text representation. Data was re-shaped to suit the numerical vectors. Scaling and weights were also experimented with but knowing that scaling primarily deals with continuous numerical data and the data was already balanced, so they didn't help much in the final model performance.

7 Baselines

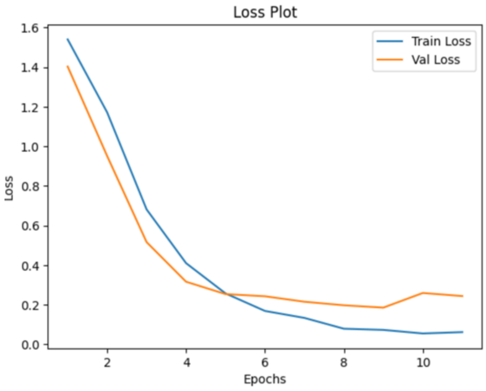
Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) are used as baselines for this task. CNN, even though widely used for image processing tasks is also able to capture local patterns and features in the text. Convolutional filters of different sizes can detect important n-gram features (word sequences or phrases) which can effectively help understanding the meaning and context of the input text. CNNs are also helpful in identifying text hierarchy distinguishing between individual words and phrases. LSTM, on the other hand, are beneficial as they can remember the earlier parts of text while analyzing newer bits which is extremely crucial to building context and understanding it based on dependencies[15]. These properties make both these models ideal choice for the BBC Classification dataset that we're dealing with.

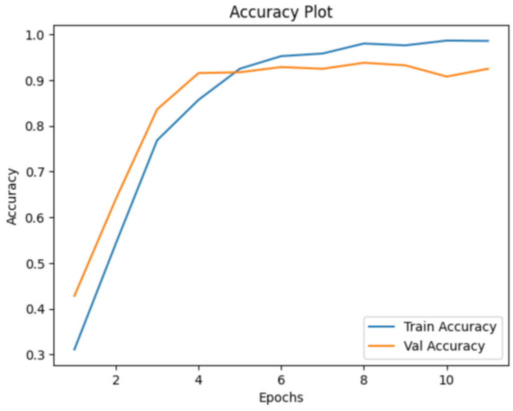
8 Results, error analysis

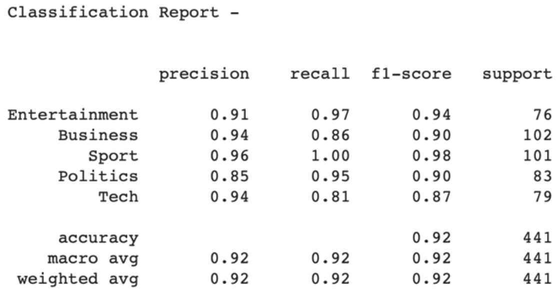
After pre-processing the dataset, we started with a baseline accuracy of 42.83% and 10.01% for CNN and LSTM models respectively. We used GloVe vectorizer with CNN and Word2Vec with LSTM. The two vectorizers were primarily chosen for their capability to understand the meaning and context of the words and phrases efficiently.

GloVe-CNN Model:

- Accuracy: The model correctly predicted the class labels for 92% of the instances in the test set.
- Precision: The model achieved high precision scores for Entertainment (91%), Business (94%), Sport (96%), Politics (85%), and Tech (94%), indicating that it made relatively few false positive predictions for these classes.
- Recall: The model achieved high recall scores for Entertainment (97%), Sport (100%), and Politics (95%). These results indicate that the model effectively identified most instances belonging to these classes.
- F1-Score: The GloVe-CNN model achieved high F1-scores for Entertainment (0.94), Business (0.90), Sport (0.98), Politics (0.90), and Tech (0.87), indicating good overall performance across these classes.





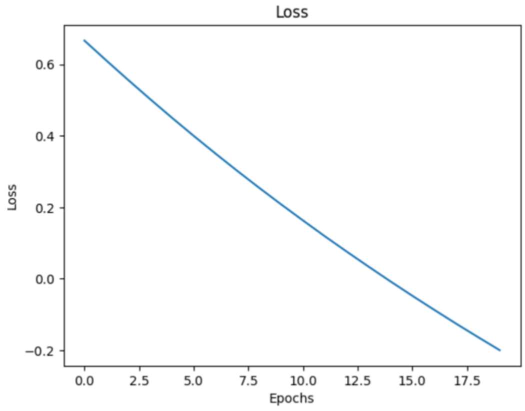


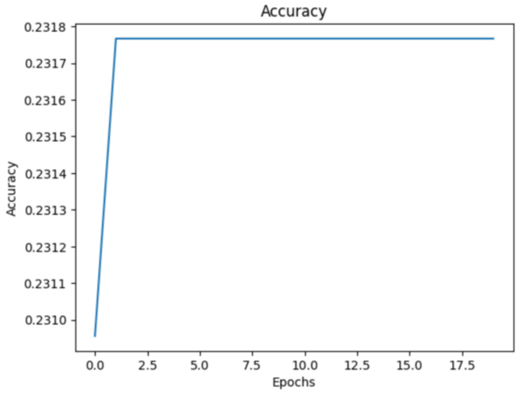
```
Classification Report -

              precision    recall  f1-score   support

Entertainment      0.91      0.97      0.94        76
     Business      0.94      0.86      0.90       102
        Sport      0.96      1.00      0.98       101
     Politics      0.85      0.95      0.90        83
         Tech      0.94      0.81      0.87        79

     accuracy                          0.92       441
    macro avg      0.92      0.92      0.92       441
 weighted avg      0.92      0.92      0.92       441
```

Word2Vec-LSTM Model:

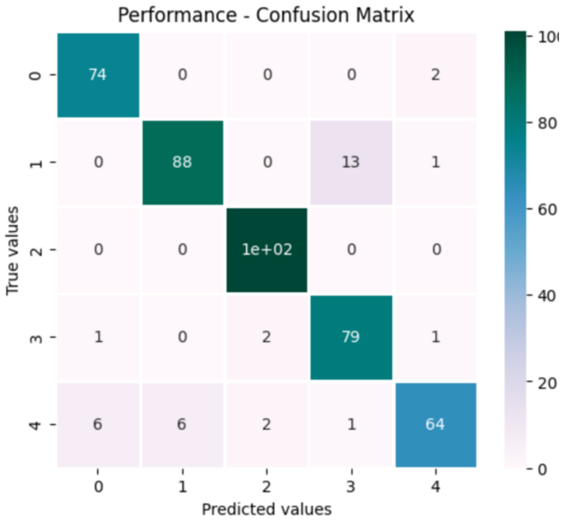
The Word2Vec-LSTM model, in contrast, performed poorly compared to the GloVe-CNN model:

- Accuracy: The model achieved an accuracy of only 17%.
- Precision, Recall, and F1-Score: The model yielded extremely low values for precision, recall, and F1-scores across all classes. This indicates that the model struggled to accurately classify instances for all categories, resulting in a poor overall performance.





The overall accuracy of 92% in GloVe-CNN's case indicates that the model generalized well to the test set whereas the poor performance of the Word2Vec-LSTM model suggests that it failed to effectively capture the semantic information and contextual dependencies within the text data. This could be due to limitations in the Word2Vec embeddings or the LSTM architecture's inability to model the complex relationships in the dataset. However, a major error that can be attributed towards the failure of this model can be erroneously using loss function for binary classification instead of categorical cross entropy loss function. By the time this error was realised, it could not be corrected due to lack of time.



Error Analysis:

Few of the snippet of the examples predicted wrong are as follows:

- Predicted: 0; True: 2
  Text: Text: moya emotional after davis cup win carlos moya describe spain 's davis cup victory as the highlight of his career after he beat andy roddick to end the usa 's challenge in seville . moya make up for miss spain 's 2000 victory through injury by beat roddick 6 - 2 7 - 6 (7 - 1) 7 - 6 (7 - 5) to give the host an unassailable 3 - 1 lead . " i have wake up so many night dream of this day , " say moya . " all my energy have be focus on today . " what i have live today i do not think i will live again . " spain 's only other davis cup title come two year ago in valencia , when they beat australia . and moya , nickname charly , admit : " the davis cup be my dream and i be a bit nervous at the outset ….

- Predicted: 0; True: 2
  Text: cole face lengthy injury lay - off aston villa 's carlton cole could be out for six week with a knee injury . the striker , who be on a season - long loan from chelsea , pick up the knock in an england under-21 match against holland early this month . " carlton will be out of action for four to six week after a bad challenge , " say villa boss david o'leary . " i will not be able to tell you whether he will need an operation until maybe next week . whether he have an operation have get to be leave to chelsea . " cole , who also struggle with an ankle problem early in the season , be unable to rest because o'leary have a shortage of striker . the return to fitness of darius vassell after four month out with a broken ankle and the emergence of luke moore have alleviate some of the villa 's manager 's problem in that department

- Predicted: 0; True: 3
  Text: burglar defence guideline issue householders who injure or even kill intruder be unlikely to be prosecute - provide they be act " honestly and instinctively " , new guideline say . the law also protect those who use " something to hand " as a weapon . the leaflet , publish by police and prosecutor , aim to combat confusion about current legislation , which let people use " reasonable force " . the guidance , relate to england and wales , follow a recent decision by minister not to change the law . do what you " honestly and instinctively " believe be necessary would be the strong evidence of act lawfully , the guidance say . and the law protect those who use " something to hand " as a weapon , say the leaflet publish jointly by the crown prosecution service (cps) and association of chief police officers (acpo) . as a general rule , the more extreme the circumstance and fear feel , …
- Predicted: 0; True: 2
  Text: blue slam blackburn over savage birmingham have confirm blackburn make a bid for robbie savage - but manage director karen brady have call it " derisory " . rover have reportedly offer £ 500,000 up front for the wales star , 30 , with the fee rise to £ 2.2 m. but brady tell the sun the bid be " a waste of fax paper and my time " . she add : " the way thing be go all this could affect the relationship between the club . they have get into robbie 's head . but he be not for sale . " savage 's future at birmingham have be the source of speculation for several week , with some fan criticise his performance for the club early in the season . however , good display against west brom and aston villa have impress blues fan . " the crowd give i a massive standing ovation when i come off on saturday which be nice…
- Predicted: 0; True: 4
  Text: humanoid robot learn how to run car - maker honda 's humanoid robot asimo have just get fast and smart . the japanese firm be a leader in develop two - legged robot and the new , improve asimo (advanced step in innovative mobility) can now run , find his way around obstacle as well as interact with people . eventually asimo could find gainful employment in home and office . " the aim be to develop a robot that can help people in their daily life , " say a honda spokesman . to get the robot run for the first time be not an easy process as it involve asimo make an accurate leap and absorb the impact of land without slip or spinning . the " run " he be now capable of be perhaps not quite up to olympic star kelly holmes ' standard…

These are just a few examples and for the sake of space issues, we're not citing all.

9 Lessons learned and conclusions

Lessons Learned:

1. In this case, the GloVe embeddings outperformed Word2Vec embeddings. It is important to carefully select embeddings that capture semantic information relevant to the task at hand.
2. The CNN architecture demonstrated superior performance compared to the LSTM architecture in this scenario. More architectures could have had been experimented with for effective comparative analysis.
3. Major lesson learnt is using the correct classifier techniques for different classifications.

Conclusions:

1. The GloVe-CNN model achieved strong performance, demonstrating its effectiveness in classifying news articles into different categories. It outperformed the Word2Vec-LSTM model in terms of accuracy, precision, recall, and F1-score.
2. The results emphasize the importance of leveraging pre-trained word embeddings, such as GloVe, which capture rich semantic information, for text classification tasks.
3. The CNN architecture proved to be more effective than the LSTM architecture for this specific dataset. It was effective in capturing local patterns and features in the text.

Future Work:

There were some potential avenues for future work:

1. Exploring ensemble methods, such as combining multiple models or using model averaging, could potentially improve the overall classification performance. Ensemble techniques have been known to enhance the robustness and generalization of models.
2. Investigating the applicability of transfer learning approaches, such as fine-tuning pre-trained language models like BERT, could be beneficial.

References

[1] Nic Newman 14th June 2023 and Newman, N. (no date) *Overview and key findings of the 2023 Digital News Report*, *Reuters Institute for the Study of Journalism*. Available at: https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023/dnr-executive-summary (Accessed: 08 May 2024).

[2] Durna, M.B. (2024) *Advanced word embeddings: Word2vec, glove, and FastText*, *Medium*. Available at: https://medium.com/@mervebdurna/advanced-word-embeddings-word2vec-glove-and-fasttext-26e546ffedbd (Accessed: 08 May 2024).

[3] Nkongolo Wa Nkongolo, M. (2023) *News classification and categorization with smart function sentiment analysis*, *International Journal of Intelligent Systems*. Available at: https://www.hindawi.com/journals/ijis/2023/1784394/ (Accessed: 08 May 2024).

[4] Hassan, A.F. and Bhaya, W.S. (no date) *Analysis of BBC News by applying classification algorithms*, *Journal of Advanced Research in Dynamic and Control Systems*. Available at: https://www.jardcs.org/abstract.php?id=3861 (Accessed: 08 May 2024).

[5] Muhammad Nabeel Asim *et al.* (no date) *A Robust Hybrid Approach for Textual Document Classification*. Available at: https://www.dfki.de/fileadmin/user_upload/import/10638_Asim_text_document_classification.pdf (Accessed: 08 May 2024).

[6] Liparas, D. *et al.* (1970) *News articles classification using random forests and weighted multimodal features*, *SpringerLink*. Available at: https://link.springer.com/chapter/10.1007/978-3-319-12979-2_6 (Accessed: 08 May 2024).

[7] *BBC News Data Classification using naïve Bayes based* ... Available at: https://www.researchgate.net/publication/360589123_BBC_NEWS_DATA_CLASSIFICATION_USING_NAIVE_BAYES_BASED_ON_BAG_OF_WORD (Accessed: 08 May 2024).

[8] Ugwuoke, U.C., Aminu, E.F. and Ekundayo, A. (2023) *Performing data augmentation experiment to enhance model accuracy: A case study of bbc news' data*, *SSRN*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4333014 (Accessed: 08 May 2024).

[9] Smalley, K. (2020) *Text classification of BBC News articles and text summarization using text rank*, *Academia.edu*. Available at: https://www.academia.edu/102138229/Text_classification_of_BBC_news_articles_and_text_summarization_using_text_rank?f_ri=288030 (Accessed: 08 May 2024).

[10] Zhou, W. (2020) *Glove.6b.100d.TXT*, *Kaggle*. Available at: https://www.kaggle.com/datasets/wzhou007/glove6b100dtxt (Accessed: 08 May 2024).

[11] Terry, A.F. (2024) *BBC Full Text Document Classification*, *Kaggle*. Available at: https://www.kaggle.com/datasets/alfathterry/bbc-full-text-document-classification (Accessed: 08 May 2024).

[12] Kushwaha, S. (2019) *BBC Full Text Document Classification*, *Kaggle*. Available at: https://www.kaggle.com/datasets/shivamkushwaha/bbc-full-text-document-classification/data (Accessed: 08 May 2024).

[13] (No date) *(PDF) practical solutions to the problem of diagonal dominance in kernel document clustering*. Available at: https://www.researchgate.net/publication/221346280_Practical_solutions_to_the_problem_of_diagonal_dominance_in_kernel_document_clustering (Accessed: 08 May 2024).

[14] Zhou, W. (2020a) *Glove.6b.100d.TXT*, *Kaggle*. Available at: https://www.kaggle.com/datasets/wzhou007/glove6b100dtxt (Accessed: 08 May 2024).

[15] Sawant, M. (2019) *Text sentiments classification with CNN and LSTM*, *Medium*. Available at: https://medium.com/@mrunal68/text-sentiments-classification-with-cnn-and-lstm-f92652bc29fd (Accessed: 08 May 2024).