# Classification of Underrepresented Text Data in an Imbalanced Dataset Using Deep Neural Network

Humaira Zahin Mauni
*Department of Computer Science and Engineering (CSE)*
*Ahsanullah University of Science and Technology*
Dhaka, Bangladesh
hzm496@gmail.com

Tajbia Hossain
*Department of Computer Science and Engineering (CSE)*
*Ahsanullah University of Science and Technology*
Dhaka, Bangladesh
tajbiahossain@gmail.com

Raqeebir Rab
*Department of Computer Science and Engineering (CSE)*
*Ahsanullah University of Science and Technology*
Dhaka, Bangladesh
jishan005@gmail.com

***Abstract*—Text classification is a well researched, much-explored topic for burgeoning researchers in the field of data mining. Yet, all the inadequacies that actual text data present during practical applications of these models in real life calls forth the need for further study. Imbalanced datasets, in particular, provide a roadblock to common high-performance algorithms. Thus, our research explores text classification concerning imbalanced datasets containing underrepresented categories, using the favored machine learning algorithms of today – neural networks. We have looked to explore how neural network classification models can improve upon inadequate datasets. Our research showcases why deep learning is the preferred classification algorithm, and proves, with hard evidence, that it outperforms other machine learning algorithms in text classification.**

***Index Terms*—artificial intelligence, deep neural network, big data for development, text classification, imbalanced data**

## I. Introduction

Artificial intelligence has revolutionized the way we monitor and evaluate large scale data for sustainable development goals. When practically implementing sustainable development strategies, automatic collection and organization of big data is a crucial step for data scientists. However, the distinction between *big data* and *big data for development* forces us to reevaluate the common strategies we apply in automated text classification using artificial intelligence. Sources of big data for development are those which can be analyzed to gain insight into human well-being and development [1], and thus the data collected cannot simply be discarded if they are underrepresented in the dataset as a whole. With massive quantities of data, there is a risk of focusing exclusively on finding patterns or correlations and subsequently rushing to judgments without a solid understanding of the deeper dynamics at play [2]. Therefore, when classifying text data with development in mind, it is important to be aware of the digital divide present between developed and developing countries, where data present may be imbalanced and underrepresented. We have observed from our findings that imbalanced data tends to be classified less accurately compared to properly represented, balanced data. Computational intelligence, specifically neural network algorithms, is the solution to overcoming this divide. The deep neural network model we have proposed presents significant improvements in this area.

For our research, we have focused on the text classification of imbalanced data, and how to improve upon the accuracy when classifying underrepresented categories. Normal machine learning algorithms give the impression of successful classifications, but upon a closer look, underrepresented categories are shown to be classified poorly and are not reflected in the overall results. The deep learning algorithm proposed seeks to eliminate this problem while maintaining acceptable overall accuracy within the results.

## II. Literature Review

In recent years, studies on the classification of text data have been mostly concerned with the application of machine learning models on balanced data. Limited research has been done on classification techniques on imbalanced data using deep learning. Selected research papers on classification of text data using machine learning and deep learning algorithms are described as follows-

Ali Selamat, Hidekazu Yanagimoto and Sigeru Omatu [3] proposed a web page classification method based on a principal component analysis (PCA) method and class profile based features (CPBF) applied to neural network classification. With their proposed methodology, they obtained 98.80% and 96.15% accuracies on two of the classes from the Yahoo sports news dataset. They have also overcome the limitation of PCA in supervised data.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao [4] introduced a Recurrent Convolutional Neural Network (RCNN) for text classification. Previous studies showed that Recurrent Neural Network (RNN) is a biased model and Convolutional Neural Network (CNN) can be used to tackle this problem. To address the limitations of previous studies an RCNN model is proposed. The RCNN model applied to four datasets gave an accuracy of 96.49% on 20Newsgroups, 95.20% on Fudan, 49.19% on ACL, and 47.21% on the SST dataset.

Matthias Damaschk, Tillmann Dönicke, and Florian Lux [5] held out the issues presented by text data when it is presented in a form not suitable to machine learning algorithms. They

978-1-7281-7366-5/20/$31.00 ©2020 IEEE

tested out different feature extraction techniques to combat these adversities and classified the lyrics using three different neural network algorithms. This paper showed us that neural network algorithms are well-preferred over other algorithms when it comes to dealing with imbalance.

In real-world applications the importance of imbalanced text classification is high. Aixin Sun, Ee Peng Lim, and Ying Liu [6] suggest that the distribution of examples across the known classes is biased in some areas. This paper presented strategies for handling imbalanced classification using the Support Vector Machine (SVM) classifier. The macro averaged imbalanced ratio of SVM on 20Newsgroups, Reuters, and WebKB dataset were 19.3, 116.4, and 24.1 respectively, showing that the classical machine learning algorithm performs poorly on imbalanced datasets.

## III. Methodology

A common approach to text data classification is the matching of news categories based on the contents of news articles. We chose to follow this convention, except with the additional task of carrying out text classification on both balanced and imbalanced datasets.

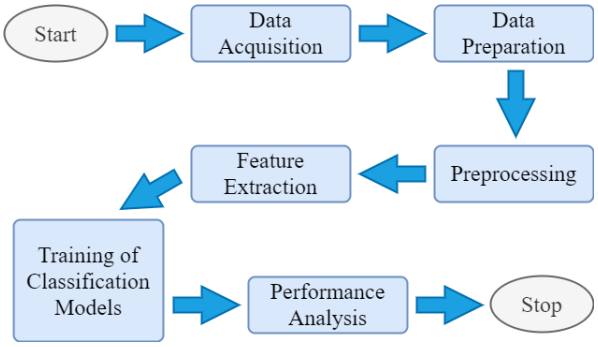


Fig. 1. Proposed Methodology

### A. Data Acquisition and Preparation

Reuters-21578, 20 Newsgroups, BBC News, and Stanford are some popular datasets mostly used in text classification. The BBC news dataset is one of the most robust, balanced, and topical amongst those mentioned [7]. There are 2225 different articles with 5 categories – which makes it an ideal dataset for the purpose of our research. After data acquisition, two different versions of this dataset was prepared -

1) **A balanced dataset**, where the news articles are distributed fairly evenly across the 5 classes.
2) **An imbalanced dataset**, where the distribution of articles is manually manipulated, ie. reduced, to create imbalance.

### B. Preprocessing

To prepare our text data for machine learning purposes, we must change it's structure to one more suitable for our classification model to analyze and extract relevant information from. An effective preprocessor represents the document efficiently in terms of both space and time requirements and maintain good retrieval performance (precision and recall) [8]. "Fig. 2" summarizes this process in 2 steps -

TABLE I
BBC News Articles Dataset Distribution

| Categories | Number of Articles | |
|---|---|---|
| | ***Balanced Dataset*** | ***Imbalanced Dataset*** |
| Business | 510 | 510 |
| Entertainment | 386 | 100 |
| Politics | 417 | 100 |
| Sports | 511 | 511 |
| Technology | 401 | 100 |

1) **Dataset Extraction** - A simple extraction algorithm is carried out to collect all the files from their respective folders and are combined into a single comma-separated value (CSV) file.
2) **Noise Removal** - We use the umbrella term *noise removal* to refer to the process of cleaning the text and removing any unnecessary information from it. This includes text cleaning (to remove punctuation marks and extra space), stopword removal (removal of all language-specific functional words, which are frequent words that carry no information i.e., pronouns, prepositions, conjunctions) [8], and lemmatization (grouping together the different inflected forms of words so they can be analyzed as a single term using the Natural Language Toolkit (NLTK) corpus [9]).

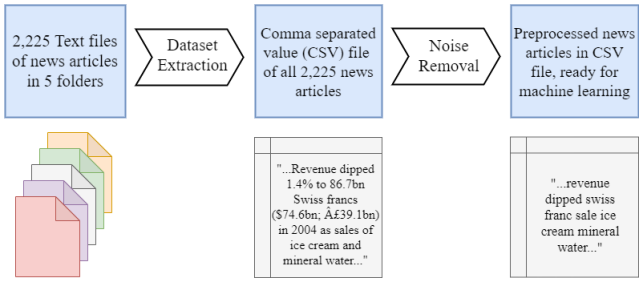


Fig. 2. Steps in data preprocessing

### C. Feature Extraction

Words of the text data are usually represented as discrete, categorical features. Feature extraction maps such text data to real-valued vectors for our classification model [10]. Extracting a set of features using effective algorithms will not only reduce dimensions of the feature space, but it will also delete redundant features for our model [11]. For this study, a total of five different feature extraction methods were needed for all the classification algorithms that were applied to the BBC news dataset.

- Tokenization using a vectorizer function and term frequency-inverse document frequency (tf-idf) were used when classic machine algorithms were applied (i.e Support vector machine, Naive Bayes, Logistic Regression, etc.).

- Bag of words (BoW), word embeddings, the GloVe algorithm, and again tf-idf were applied for neural network models.

### D. Text Classification on Balanced and Imbalanced Data

We have prepared our BBC news article dataset and carried out some common classification algorithms to observe the results. These classification models include-

- Classic Algorithms - Decision Tree, Random Forest, K-Nearest Neighbor, Naive Bayes, Support Vector Machine, and Logistic Regression classifier.
- Neural Network Algorithms - Recurrent Neural Network (RNN) and variants of Deep Neural Network (DNN) algorithms with different feature extraction methods.

From the classic algorithms, **Logistic Regression** gave us the highest accuracy out of those tested on our chosen dataset and was thus chosen as our baseline for comparison. From the neural network algorithms, **Deep Neural Network** using bag of words had the highest accuracy out of all the models, giving us our proposed model. Thus, instead of comparing the results across each class for many models, we obtained the two best to observe how to minimize the effect of imbalance.

**Baseline Model - Logistic Regression**

Logistic Regression is a predictive analysis algorithm and based on the concept of probability. It assigns observations to a discrete set of classes [12]. This is our baseline model for comparison.

**Proposed Model - Deep Neural Network**

A Deep Neural Network is a neural network with more than two layers, which uses sophisticated mathematical modeling to process data in complex ways [13]. This was our proposed model.

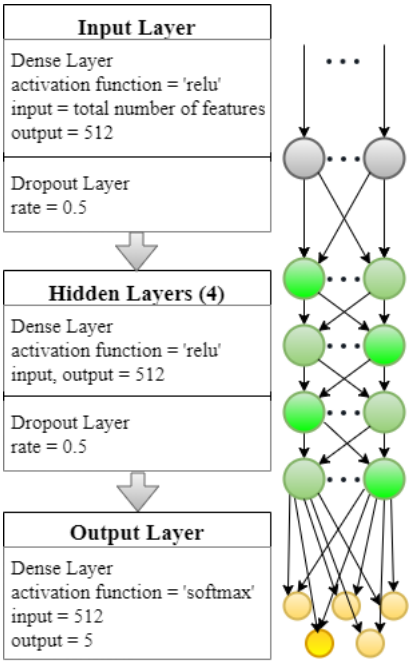


Fig. 3. Proposed Deep Neural Network Model

DNNs have interconnected layers where every single layer receives information from the previous one, and in return sends information to the next. The input layer in a DNN is a connection of feature space which could be tf-idf, word embeddings, etc. The output layer houses neurons equal to the number of classes. The multiclass DNN model we have created uses the basic bag of words approach for the input layer, standard backpropagation algorithm and 'ReLu' as the activation function for the hidden layers, and 'Softmax' function for the output layer.

Thus, these two text classification models were carried out on both our balanced and imbalanced dataset, and the result was obtained for analysis.

## IV. Performance Analysis

When dealing with imbalance, high accuracy can be obtained while ignoring other useful performance indicators. This is because accuracy is a good metric only when class labels have even distribution. Thus, we turn to other evaluation metrics alongside accuracy.

### A. Evaluation Metrics

1) Precision - Precision is the ratio between correctly predicted positive classifications to the total predicted positive classifications.
2) Recall - Recall is the ratio between correctly predicted positive classifications to all positive and negative in the category.
3) F1 Score - Used to measure classification when there is an uneven number of observations in each class, it is the harmonic mean of precision and recall.
4) Macro averaged F1 score versus accuracy - Macro averaged F1 score is the average of each category's F1 scores, without taking into consideration the number of samples in each class. When classifying imbalanced datasets, we hope to achieve similar accuracy and macro-averaged F1 score, so that the accuracy reflects the classification for all classes.

### B. Results

*1) Classification of Balanced Dataset:* The result of our research in "table.II" indicated that classic algorithms are outperformed by our deep neural network algorithm (98.65% accuracy) when classifying balanced text data. Comparing the F1 scores of each category present in the dataset gives us an idea of how accurately each category is classified. We can see from the table below that both algorithms perform outstandingly when classifying a balanced dataset, and all categories have similar F1 scores. The macro averaged F1 scores and the accuracies do not differ by much, proving that the accuracy reflects the correct classification for each class.

*2) Classification of Imbalanced Dataset:* However, when we apply the same algorithms on our imbalanced dataset- where some categories have a reduced number of articles-the performance degrades for logistic regression.

TABLE II
RESULTS OF LOGISTIC REGRESSION AND DEEP NEURAL NETWORK APPLIED ON A BALANCED DATASET

| Categories | F1 Score | |
|---|---|---|
| | ***Logistic Regression*** | ***Deep Neural Network*** |
| Business | 97% | 98% |
| Entertainment | 99% | 99% |
| Politics | 97% | 97% |
| Sports | 99% | 99% |
| Technology | 98% | 100% |
| **Accuracy** | 97.75% | 98.65% |
| **Macro Avg F1 Score** | 98% | 99% |

TABLE III
RESULTS OF LOGISTIC REGRESSION AND DEEP NEURAL NETWORK APPLIED ON AN IMBALANCED DATASET

| Categories | F1 Score | |
|---|---|---|
| | ***Logistic Regression*** | ***Deep Neural Network*** |
| Business | 94% | 98% |
| [a]Entertainment | 91% | 92% |
| [a]Politics | 81% | 98% |
| Sports | 99% | 100% |
| [a]Technology | 90% | 91% |
| **Accuracy** | 94.33% | 97.74% |
| **Macro Avg F1 Score** | 91% | 96% |

[a]Reduced classes

In "table.III", the overall accuracy shows that our deep neural network algorithm came out on top. However, our real problem lies in the classification of our underrepresented categories - the three reduced classes being entertainment, politics, and technology.

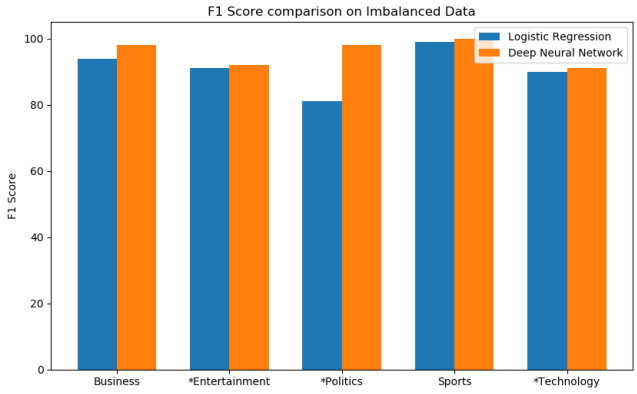


Fig. 4. Comparison between Logistic Regression and DNN on an Imbalanced Dataset

When comparing the F1 scores for the 3 target categories for under-representation, we can observe how the deep neural network algorithm performs. Our proposed model had a higher F1 score for all 3 categories compared to the logistic regression model. We also achieved a higher F1 score on the imbalanced dataset over the balanced dataset in one category (politics). The most conclusive proof is the difference in accuracy and macro-averaged F1 scores - DNN's accuracy reflected the classification of the imbalanced classes more concisely than logistic regression.

## V. CONCLUSION AND FUTURE WORK

In conclusion, our research has proved that text classification on data for development benefits the most from deep learning, as deep neural network algorithms outperform commonly preferred methods when it comes to dealing with imbalanced data. The difference in accuracy and macro averaged F1 score is minimized, and overall F1 scores and accuracy are also higher. Thus, we can conclude that deep neural network models require further research and show promise in big data for development, specifically concerning sustainable development goals. The promise of Big Data for Development is and will be, best fulfilled when its limitations, biases and ultimately features, are adequately understood and taken into account when interpreting the data [1].

In the future, the possible next step is researching the effect of deep neural networks on real-life datasets obtained for sustainable development goals. Further work in the area includes applying transfer learning, different feature extraction methods compatible with deep neural networks (like PCA), and other variants of neural network models.

## REFERENCES


[1] "Big data and global development," https://www.sas.com /en_us/insights /articles/big-data/big-data-global-development.html (accessed: 28 February, 2020).

[2] "Big Data for Development: Challenges and Opportunities," UN Global Pulse, May 2012. [Online]. Available: https://beta.unglobalpulse.org/wp-content/uploads/2012/05/BigDataforDevelopment-UNGlobalPulseMay 2012.pdf.

[3] A. Selamat, H. Yanagimoto, and S. Omatu, "Web news classification using neural networks based on PCA," in Proceedings of the 41st SICE Annual Conference. SICE 2002.,vol. 4, pp. 2389–2394 vol.4, Aug 2002.

[4] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in Proceedings of the AAAI Conference on Artificial Intelligence, 2015.

[5] T. Donicke, F. Lux, and M. Damaschk, "Multiclass text classification on unbalanced, sparse and noisy data," in Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing, p. 58–65, 09 2019.

[6] A. Sun, E.-P. Lim, and Y. Liu, "On strategies for imbalanced text classification using SVM: A comparative study," Decision Support Systems, vol. 48, p. 191–201, Dec. 2009.

[7] D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", in Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, Pennsylvania, USA, 2006.

[8] V. Srividhya and R. Anitha, "Evaluating preprocessing techniques in text categorization," International Journal of Computer Science and Application, Issue 2010.

[9] "Text Preprocessing in Python: Steps, Tools, and Examples," https:// medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908 (accessed: 28 February, 2020).

[10] "Machine Learning-Text Processing," https://towardsdatascience.com/ machine-learning-text-processing-1d5a2d638958 (accessed: 25 February, 2020).

[11] H. Liang, X. Sun, Y. Sun, Y. Gao, "Text feature extraction based on deep learning: a review", EURASIP Journal on Wireless Communications and Networking, 211 (2017).

[12] "Logistic Regression — Detailed Overview," https://towardsdatascience .com/logistic-regression-detailed-overview-46c4da4303bc (accessed: 8 January, 2020).

[13] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes and D. Brown, "Text Classification Algorithms: A Survey", Information, vol. 10, no. 4, p. 150, 2019.