Performing Data Augmentation Experiment to Enhance Model Accuracy: A Case Study of BBC News' Data

Uchenna Cosmas Ugwuoke[1], Enesi Femi Aminu[2,*] and Ayobam Ekundayo[3]

[1,2,3]Department of Computer Science, School of Information & Communication Technology, Federal University of Technology, Minna Nigeria

*Corresponding author

Abstract.

In natural language processing, text classification forms an essential task to be performed; as such, the use of machine learning algorithms have constantly become indispensable and significance to the research drive. However, the problem of solving text classification with the traditional models gets more challenging because of ambiguities associated with natural languages. A typical example is synonyms' concept mismatch, and other related issues that accurately attribute text to their related contexts. While a more robust model with an increased number of hidden layers such as LSTM is essential, because of the volume of data involved; exploration of strategies for data augmentation is highly significant. To this end, this research aims to employs semantic lexical database, called WordNet as strategy to augment the BBC news textual data obtained from kaggle repository. This is to pave way for a more efficient news data classification based on the proposed LSTM model. The total BBC news samples are 2,225 data points, and each data point is grouped into five different news categories, which include, technology news, business news, sport news, entertainment news, and political news. Experimental evaluations are carried out using the benchmark BBC news dataset; and the newly augmented dataset within the scope of this study. Consequently, the accuracy of the classification LSTM model for original news dataset and the augmented dataset are 90% and 95% respectively. Therefore, the proposed data augmentation strategy is promising for textual datasets.

Keywords: Data augmentation, WordNet, BBC news data, LSTM model, contextual meaning

1. Introduction

Presently, the application of artificial intelligence (AI) technique to any real life domain has significantly increased due to its overwhelming performance in providing solutions to complex problems. Conversely, the performance of deep learning models outweigh the classical machine learning techniques (Chlap et al., 2021). However, for an optimal performance of the (deep) models, it requires a larger volume of data for training. But the availability of huge volume of datasets in most cases are serious issues to contend with by researchers because of limited datasets (Duong & Nguyen-Thi, 2021). This is because to obtain a well annotated datasets, a lot of resources (in terms of time and costs) are involved. Thus, the need for data augmentation technique becomes inevitable to deal with the

challenges of data sparsity during the process of text classification. Also, the techniques have severally served as booster for the accuracy of learning models (Hernandez-Garcıa & Konig, 2018).

It is an undisputable fact that during the last decade, text classification, an indispensable task in natural language processing has witnessed constant attention from researchers as a result of the potentiality of both traditional machine learning and deep learning algorithms. Many of these models have been consistently employed to classify text, (news data classification) for example. Naïve Bayes (NB) is the first classification technique employed for the task of text classification (Li, et al., 2020). Afterward, other traditional models otherwise known as classifiers, such as Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbour (KNN) followed. Although, in a recent time eXtreme Gradient Boosting (XGBoost) has arguably identified to possess the capability to produce far reaching accuracy as compared to the previous classifiers.

However, as the problem of solving text classification with the traditional models get more challenging; a more robust models with an increased numbers of hidden layers such as Long Short Term Memory (LSTM), Bidirectional Encoder Representation from Transformers (BERT), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) offer more potentials to deal with the problem (Dogru, et al., 2021; Joshi, et al., 2020). Although, some of the models (CNN) are not primarily designed for textual data classification. Therefore, news data classification such as the BBC textual sample dataset obtained from kaggle online repository in this research equally requires the techniques of data augmentation as a result of ambiguity in natural languages. Consequently, effective news classification strategy has to be deployed to obtain contextual meaning of news article. However, only limited numbers of research have been carried out in this regard.

Furthermore, as earlier stated, it is unarguable fact that the efficiency of the top notch models (such as LSTM) heavily relies on huge volume of datasets. Besides, the availability of all forms of data are actually scarce or limited (Shorten & Khoshgoftaar, 2019); although depends on the domain for the dataset. Consequently, the technique of data augmentation is promising to provide the needed palliative to cushion the efficacy of the deep learning models. However, based on literature, images type of data augmentation strategies have enjoyed better attention from the researchers than the textual forms. One of the fundamental reasons attributed to text based augmentation algorithms is the lack of generalized rules for natural language transformation owing to the ambiguity of natural language itself (Wei & Zou, 2019). Therefore, interested researchers resulted to devise a mechanism to augment textual data by translating sentences from one language to another (Yu et al., 2018). Other devised mechanisms used include data noising as smoothing (Xie et al., 2017); and predictive language mechanisms for replacement of synonym (Kobayashi, 2018).

Similarly, beyond data augmentation that is typically used in image classification by generating more image data, the technique is equally applied for textual data classification by also generating more data but leverage on different strategies (Duong & Nguyen-Thi, 2021). This is because data augmentation is more complex in textual form of data. More so, an improvement of original datasets either to increase the size, or process the relevant data in terms of contextual meaning or feature extraction via the strategy of data augmentation is

not limited to a particular domain but across different real life scenarios. It cut across different spheres of domains such as biomedical (Geng et al., 2022); computational chemistry (Ulrich et al., 2021).

One of the primary strengths of data augmentation techniques largely lies on its ability to prevent overfitting. This is achieved by adjusting or amending the limited datasets to possess the qualities of big data (Shorten & Khoshgoftaar, 2019). Necessary precaution has to be observed when choosing the strategy of data augmentation because there is tendency of overfitting or poor performance during training when increased volume of augmented data are considered (Duong & Nguyen-Thi, 2021). This scenario may occur if the augmented data are closely similar or different from the original data. In view of this development, this research aims to exploit a semantic lexical database called WordNet (Chakravarthi et al., 2018) to augment the BBC dataset for the sole goal of contextual meaning of the data based on LSTM model. WordNet, a semantic lexical data source similar in structure as ontology has over time prove its capacity to close the gap of ambiguities associated with the natural languages (Jarrar, 2021). WordNet typically consist of three databases namely, nouns subnet, verbs subnet and adjective/adverb subnet databases in support with semantic relations; such as synonyms, hypernyms, hyponyms, holonymys, and meronyms (Fawei et al., 2019; Uthayan & Mala, 2015). An experimental evaluation would be conducted to compare the results of LSTM model without augmentation strategy and LSTM model with WordNet data augmentation strategy. Finally, the remaining sections of this paper are organized as follows: section 2 accounts for the related work while, the proposed methodology is presented in section 3. Section 4 presents the discussion of the results based on the methodology; and section 5 presents the conclusion of the paper. STOP

2. Related Work

This section gives a brief related literature on data augmentation techniques on textual and images data, but with a close focus on the former. Expectedly, the techniques are not considered in isolation that is, the application of both machine learning and deep learning models for text classification are equally considered in this section. To this end, the literature of (Wei & Zou, 2019) aims to improve the performance of CNN-RNN based text classification model using easy data augmentation, which consist of random swap, random insertion, random deletion and synonym replacement. The researchers admitted that for efficiency use of EDA, a synonym dictionary (WordNet) would have improved its accuracy. However, they encountered challenge of computational accessibility. Similarly, the research of (Duong & Nguyen-Thi, 2021) acknowledged the significance of data augmentation for small dataset. Easy Data Augmentation (EDA) strategy that was first proposed by (Wei & Zou, 2019) was employed for the Vietnamese data. The objective of the technique is solely to obtained synonym replacement randomly in a given sentence of dataset based on logistic regression (LR) and support vector machine (SVM). However, the researchers admitted that the approach is deficient because it can increases the irrelevant sentences.

Furthermore, there are various techniques available for data augmentation. The survey work of (Feng et al., 2021) accounted for some of these techniques considered for natural language processing. They include EDA, feature space, as rule based technique; other techniques are model based techniques, and mixed sample data augmentation otherwise

known as interpolation technique. Also, there is always a difficulty in the choice of appropriate data augmentation techniques consequently, experimental evaluation becomes inevitable to untie the quagmire of effective technique to choose. This is the motivation behind the work of (Porcu et al., 2020) who conducted an evaluation to compare the accuracy on the emotion recognition of the facial expression recognition systems. Geometric transformation and oversampling augmentations were considered on the CNN deep network specifically, the VGG16 based on the given datasets. However, generative adversarial networks (GAN) is observed to be deficient in relevant computational complexity.

Data augmentation approach was also reported to be used in the literature of (Shoaib et al., 2022) as a result of limited data available. The work whose aim was to classified brain tumors based on four CNN architectures got better accuracy with the augmented dataset. However, as a result of the data augmentation strategy adopted, the size of the dataset is still small, which inadvertently affects the accuracy of the deep networks based models. Besides, it is deficient in computational complexity as a result of the data size. More so, owing to the importance of classification techniques, Ahmed and Ahmed (2021) carried out a review work on some of the existing techniques or classifiers for categorizing digital news articles. The classifiers considered were Naïve Baiyes (NB), K-Nearest Neighbors (KNN), Logistic Regression (LR) and Support Vector Machine (SVM). The result of the experiment adjudged NB to be the best among the four classifiers; the rationale behind its performance is hinges on different sources of datasets. However, the researchers suggested that more accuracy can be obtained if word or term similarity ranking technique is adopted or adapted especially on a more volume of datasets.

More importantly, text data can emanate from different sources, which include news data, chats, web data, emails, social media, user reviews, to mention a few. No doubt, text is an extremely rich source of information. However, to classify or extract insights from text can be laborious and time-consuming, as a result of its unstructured nature; thus, classification techniques are required. According to the work of Minaee, et al., (2021), text classification can be performed either through manual annotation or by automatic technique. But with the rising scale of text data in industrial usage, automatic text classification is becoming increasingly significant. Automatic text classification techniques are categorized into two approaches. They are rule-based methods and machine learning (data-driven) based methods. More so, the volume of unstructured data in the digital space call for research attention in this present age. This development motivated the research of Dogru, et al., (2021) to advanced investigation on a reliable automatic text classification. On this note, the researchers employed the traditional machine learning classifiers such as Support Vector Machine, Gauss Naive Bayes, Random Forest, and Naive Bayes and the deep learning algorithm: CNN on two English news datasets. The result shown that the deep learning technique outperformed all the other four machine learning techniques on the two datasets. However, the specified value of Doc2Vec method as word embedding approach is not based on any known model.

Similarly, the work of Ramdhani, et al., (2020) canvasses the capacity of deep learning in handling text mining for natural language processing. This is why the researchers were motivated to employ the use of Convolutional Neural Network (CNN) for classification of Indonesian language. The experiment was carried out on four hundred and seventy two

(472) Indonesian based news text data from different sources on four major classifications. These are technology, news headlines, entertainment, and sports. While the researchers hail CNN for its accuracy, they however, suggested for more datasets; and more importantly, an improvement by developing variation in the hidden layers. Besides, they suggested further that a different activation function aside sigmoid; and a different deep learning algorithm aside the CNN could be considered. In conclusion, no natural language processing's technique or algorithm is considered in the review work to handle the issue of natural language ambiguities for better accuracy.

Therefore, this research propose to employ WordNet, a semantic lexical database similar in structure with ontology as technique to perform data augmentation to infer the contextual meaning of the given dataset. The BBC dataset, which is 2,225 data points is obtained from kaggle repository.

3. The Methodology

The multi-steps data augmentation-modeling (DAM) approach is the proposed methodology to achieve the objective of this research. The methodology consists of five phases, which include data import, data exploration, data preprocessing, data augmentation using WordNet, and model training with the enhanced dataset as represented by the framework of Figure 1.

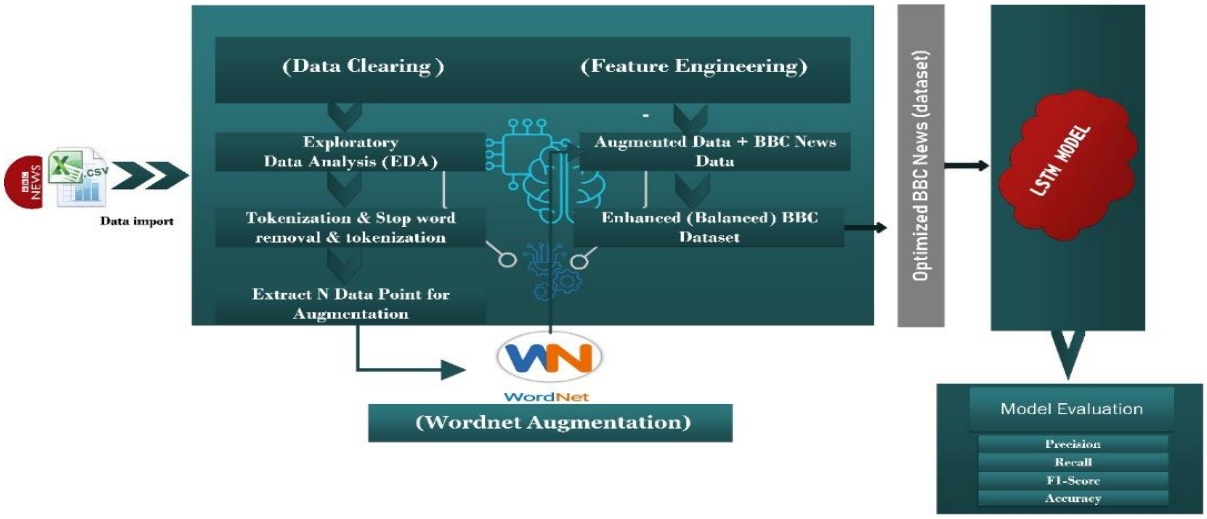

Figure 1: Conceptual Framework for the proposed Technique

From the left hand side of the framework of Figure 1, the initial phase of the methodology is to import the BBC csv dataset obtained from kaggle repository into google collaboratory environment for data science operation. The BBC news dataset contain one input feature (new content) and the output feature (the target class). The total BBC news samples is 2,225 data points, and each data point is grouped into five different classes or news group. This include; technology news, business news, sport news, entertainment news, and political news. The goal is to efficiently classify news articles into the news groups. Based on the statistical data after exploration, dataset's imbalance was identified. While sport and business news classes contain 511 and 510 data points respectively, entertainment news

class contains the least data points of 386. Technology and politics news classes contain 411 and 417 data points respectively.

The significance of preprocessing stage otherwise known as data cleaning cannot be over stated because not every content of natural language's text is relevant. If the needful is not done, it can automatically affects the results of training models. The phase include the tokenization of document, removal of punctuation marks and stop words, and stemming of candidate terms. A clear example of result of this phase is <*tv, future, hand, viewer, home, theatr*> as shown in Figure 2. That is, after stop words (such as *in, the, of*) and punctuations (such as *@, $, ^*) have been removed.

```
[ ]
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
st_words = stopwords.words('english')
from nltk.stem import PorterStemmer
import string
ps = PorterStemmer()

def dataset_cleaning(text):
    #  removeing punctuation from the dataset
    text = ''.join([token for token in text if not token in string.punctuation])
    # tokenizing the dataset e.g ['tv' , 'future' , 'in' , 'the']
    token = word_tokenize(text)
    # removing stopword (words that had less meaning to the dataset)
    text  = [ps.stem(t) for t in token if not t in st_words ]
    return text

bbc_data['clean_text'] = bbc_data["text"].apply(lambda x: dataset_cleaning(x))
```

```
bbc_data.head()
```

	category	text	clean_text
0	tech	tv future in the hands of viewers with home th...	[tv, futur, hand, viewer, home, theatr, system...
1	business	worldcom boss left books alone former worldc...	[worldcom, boss, left, book, alon, former, wor...
2	sport	tigers wary of farrell gamble leicester say ...	[tiger, wari, farrel, gambl, leicest, say, rus...
3	sport	yeading face newcastle in fa cup premiership s...	[yead, face, newcastl, fa, cup, premiership, s...
4	entertainment	ocean s twelve raids box office ocean s twelve...	[ocean, twelv, raid, box, offic, ocean, twelv,...

Figure 2: Data Preprocessing Phase

The Preprocessed data and cleaned text is represented with the feature name 'clean_text' as shown by the Figure 2. Further preprocessing includes converting the classes (technology, business, sport, entertainment, and political) into number representation (technology = 0, business = 2). This is done or achieved using Label-Encoding. Additionally, for any deep learning classification the following steps are critically important to maintain consistencies among data. They are tokenizer and word index, sequencing, and padding.

Lastly, WordNet Synsets was employed to augment the dataset to improve the prediction accuracy. That is, words synonyms and part of speech tagging were used to substitute words for generating augmented data point; and merge the augmented data point to the main BBC data point. This process is represented by the algorithmic design of Table 1.

Table 1: Algorithm Design for BBC' Dataset Augmentatio

***Algorithm3.1**: BBC Dataset Augmentation Algorithm*
***Input**: CSV Based Dataset (News Text)*
***Output**: Returns augmented datapoint*
***Parameters:** Python Packages and Classes such as edu.stanford.nlp.tagger.maxent…., py.io.BufferedReader, py.io.FileReader (list); input news text (t); tokenize(); stopWordsFxn(); wordNet; synsets(); List_of_Tokens(lt); List_of_News (ln); candidateTerms(cT)*

Procedure:

```
1    Input t
2    lt ← Tokenize (t)
3    ln ← stopWordsFxn (lt)
4    cT ← wordIndexFxn(ln);
5    Similar_concept[ ] ← wordNet(cT)   // see breakdown from step 6 to 9
6    for All cT in List_of_News :
7        Synset ← invoke wordNet [ ]
8        Similar_Concept ← Append synonyms to cT
9        Return augmented datapoint
10   End
```

From the algorithm in Table 1, the sample BBC news dataset is inputted randomly where tokenization function acts on it to produce a data structure, in this case called list of tokens (*lt*) as shown by lines 1 to 2. Consequently, a preprocessing function is invoked so as to remove stop words; the result produced is referred to as list of news (*ln*) as presented by line 3. However, there is need to still further remove unwanted terms from the list. Thus, a user defined method, *wordIndexFxnI*(*ln*) is used to perform word ranking and select the most ranking words, which constitute the candidate words (*cT*) as indicated by line 4. Lines 5 to 8 essentially introduce the use of the lexical database (WordNet) to perform the text augmentation in terms of contextual meaning to enrich the raw dataset. Finally, the similar terms which are the augmented data are appended to the initial concepts from the original dataset, which in turn serves as the input data for the LSTM based model training. The algorithm was implemented on python environment.

4. Results and Discussion

The algorithm is implemented using python 3.9.0 and the model experiment was performed in two folds; that is, LSTM model without WordNet, and with WordNet. Both models experiment were placed under the same configuration. For examples, number of embedding network layer, number of LSTM layer, number of dense layer, and same activation function was considered. Others include same loss, optimizer and metrics were used all through; lastly, the same 8 epochs (Number of training iteration) was used.

Firstly, an experiment of the LSTM benchmark training model was performed without WordNet and the result is shown by Figure 3.

```
[ ]  1 from sklearn import metrics
     2 import numpy as np
     3
     4 result = lstm_model.predict(test_pad)
     5 prediction = [np.argmax(res) for res in result]
     6 # prediction
     7 classs_report = metrics.classification_report(testy, prediction)
     8 print(classs_report)
     9
```

```
                  precision    recall  f1-score   support

               0       0.82      0.96      0.88       106
               1       0.89      0.85      0.87        66
               2       0.96      0.74      0.84        86
               3       0.99      1.00      1.00       101
               4       0.90      0.92      0.91        86

        accuracy                           0.90       445
       macro avg       0.91      0.89      0.90       445
    weighted avg       0.91      0.90      0.90       445
```

Figure 3: LSTM Benchmark Evaluation without WordNet.

The LSTM model evaluation reported an accuracy of 90% after eight epochs. The results of precision, recall and f1-score are 82%, 96% and 88% respectively.

On the other hand, the algorithm in Table 1 was fully implemented to provide data augmentation to the original dataset before training the model. As earlier explained, random sample was selected from the BBC news items; after preprocessing and tagging of the text, candidate terms was produced. Synsets and part of speech tagging were used to substitute words for generating augmented data points. Lastly, the augmented data points were merged with the original data points. Figure 4 shows the function call to WordNet routing to generate the augmented data points.

```
[ ]  1 from textaugment import Wordnet
     2
     3 wdn_aug = Wordnet(v=False ,n=True, p=0.5)
     4 # wdn_aug.augment('In the afternoon, John is going to town')
     5 # wdn_aug
```

```
'in the afternoon, trick is going to town'
```

```
[ ]  1 # def generate_aurgdata(dataset ,input_col_name, output_col_name , aug_obj):
     2 import numpy as np
     3 wdn_aud_data = generate_aurgdata(bbc_data, 'text', 'category', wdn_aug)
```

```
for [tech] : we are generating [110] arg data
for [business] : we are generating [1] arg data
for [sport] : we are generating [0] arg data
for [entertainment] : we are generating [125] arg data
for [politics] : we are generating [94] arg data
```

Figure 4: Augmented Data with WordNet

As clearly shown by Figure 4, the user defined function, *wdn_aug* of the wordnet receives a sample data to augment the five categories of the dataset. Through this function, augmented data of 110, 125, and 94 were generated for technology, entertainment and politics classes respectively. However, no data point was generated for sport class and only 1 data point was generated for business class because they both have maximum numbers of 511 and 510 data points respectively. This strategy is to ensure class balancing for effectiveness of results. A sample of the augmented data for technology class is shown by Figure 5.

```
28 # print(len(df[df[ 'category' ]== 'tech' ]))
```

Execution completed.....

	category	text
0	tech	web pic storage market hots up an increasing i...
1	tech	nintendo adds media playing to ds nintendo is ...
2	tech	digital guru floats sub-$100 pc nicholas negro...
3	tech	microsoft unveiling security peter microsoft i...
4	tech	us peer-to-peer pirates convicted the first co...

Figure 5: Sample of Technology Data Generated by WordNet

This is how the proposed WordNet generated augmented data for the categories of the news dataset with minimum data points. Consequently, the new augmented dataset was fed into the model with same configuration, an improved accuracy of 95% is obtained as shown by Figure 6.

```
[ ]  1 wd_lstm_model.evaluate(stest_pad, stesty)

16/16 [==============================] - 1s 74ms/step - loss: 0.1561 - accuracy
[0.15608465671539307, 0.951076328754425]
```

```
1 from sklearn import metrics
2 import numpy as np
3
4 result = wd_lstm_model.predict(stest_pad)
5 prediction = [np.argmax(res) for res in result]
6 # prediction
7 classs_report = metrics.classification_report(stesty, prediction)
8 print(classs_report)
```

```
              precision    recall  f1-score   support

           0       0.81      0.88      0.84        43
           1       0.99      0.96      0.97       150
           2       0.95      0.91      0.93       129
           3       0.93      1.00      0.96        37
           4       0.97      0.98      0.97       152

    accuracy                           0.95       511
   macro avg       0.93      0.95      0.94       511
weighted avg       0.95      0.95      0.95       511
```

Figure 6: LSTM + WordNet Benchmark Evaluation

Figure 6 depicts the results of how augmented data were generated using WordNet Synsets to enhance the dataset for better prediction accuracy. However, the results of precision and recall tends to fluctuate although, depending on the iteration. The experiment

is still work in progress because the researchers are currently employing word embedding algorithm (word2vec) to equally perform data augmentation experiment. Table 2 presents the results comparison of the model with and without WordNet.

Table 2: Comparison of Results Performance

Model/Data Aug. Strategy	Accuracy (%)	F1-score (%)
LSTM/None	90	87, 100, 91 (For iterations 1, 3 and 4)
LSTM+WordNet	95	97, 96, 97 For iterations 1, 3 and 4

Evidently, the proposed LSTM model with WordNet performs better than the ordinary LSTM model without data augmentation strategy as represented by the graph of Figure 7. The results of f-measure fluctuate as iteration 3 performs better than the proposed model; however, iterations 1 and 4 of the proposed model perform better.

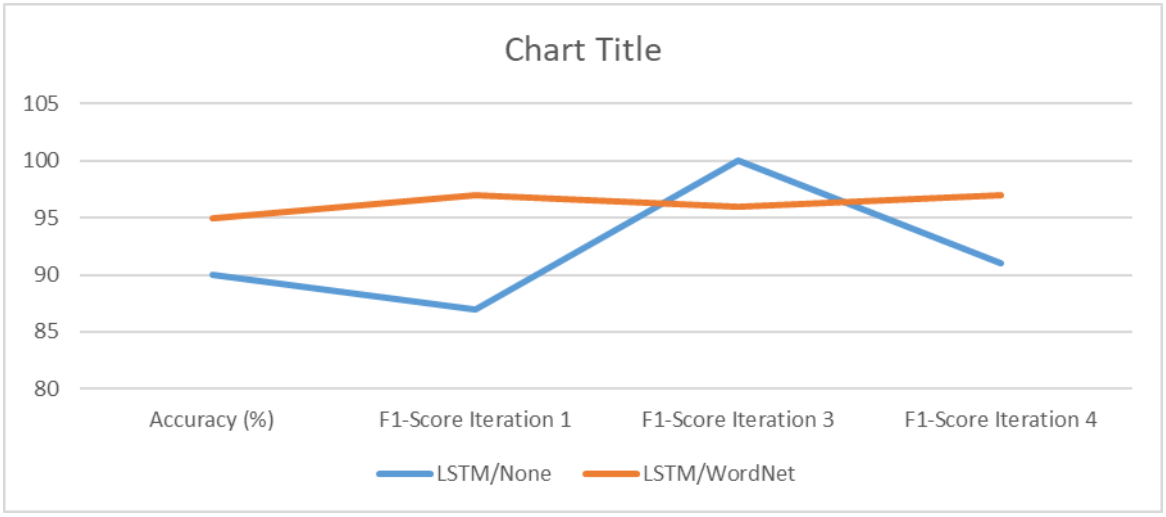

Figure 7: Graphical Comparison of Results

Figure 7 clearly presents the results comparison as depicted earlier by Table 2.

5. Conclusion

Data augmentation is an essential component of text classification in natural language processing. It has constantly demonstrates capacity to improve accuracy of models. However, based on literature some of the technique for data augmentation have shown superiority over others whether textual or images data. Therefore, there is need to perform experiment on any choice of technique to augment a given set of data. In lieu of this development, this research aims to perform experiment and ascertain the viability of WordNet (a semantic lexical data source) to augment BBC news dataset. The dataset, which consist of 2,225 data points was obtained from kaggle repository. Considering the size of the data, this research used the LSTM to design the news classification model. The news is classified into five categories, which include technology news, sports news, entertainment news, business news and political news. Series of preprocessing techniques such as removal of stop words and the user defined word tagging method were employed. At the end, a

semantic based algorithm for data augmentation is proposed in this research. The proposed LSTM with WordNet model turns out an accuracy of 95% against 90% for LSTM model without any strategy for data augmentation.

Even though, WordNet has brought a significant improvement of accuracy on the same benchmark dataset, there is need to equally improve on the consistencies of precision and recall's results, which is a function of f1-score as they tends to fluctuate among the iterations. Meanwhile, this research is still work in progress as the researchers are currently using word2vec to ascertain the superiority of the two strategies. More so, in future the researchers hope to exploit other deep learning algorithms such as BERT with Attention mechanism to validate the strength of LSTM with and without data augmentation technique. Besides, Wikipedia data source may be more efficient to deal with compound words or phrases in a given dataset; and at the overall, related synonymous words in contextual meaning are expected to be produced.

References

Ahmed, J., & Ahmed, M. (2021). Online News Classification Using Machine Learning Techniques. *Iium Engineering Journal*, *22*(2), 210-225.

Chakravarthi, B. R., Arcan, M., & McCrae, J. P. (2018). Improving wordnets for under-resourced languages using machine translation. *GWC 2018 - 9th Global WordNet Conference*, *2018-Janua*, 77–86.

Chlap, P., Min, H., Haworth, A., Vandenberg, N., & Dowling, J. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, *65*, 545–563. https://doi.org/10.1111/1754-9485.13261

Dogru, H. B., Tilki, S., Jamil, A., & Hameed, A. A. (2021). Deep Learning-Based Classification of News Texts Using Doc2Vec Model. In *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)* (pp. 91-96). IEEE.

Duong, H.-T., & Nguyen-Thi, T.-A. (2021). A review : preprocessing techniques and data augmentation for sentiment analysis. *Computational Social Networks*, *8*(1), 1–16. https://doi.org/10.1186/s40649-020-00080-x

Fawei, B., Pan, J. Z., Kollingbaum, M., & Wyner, A. Z. (2019). A Semi-automated Ontology Construction for Legal Question Answering. *New Generation Computing*, *37*(4), 453–478. https://doi.org/10.1007/s00354-019-00070-2

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A Survey of Data Augmentation Approaches for NLP. *ArXiv Preprint ArXiv:2105.03075.*

Geng, M., Xie, X., Liu, S., Yu, J., Hu, S., Liu, X., & Meng, H. (2022). Investigation of data augmentation techniques for disordered Speech Recognition. *ArXiv Preprint ArXiv:2201. 05562*.

Hernandez-Garcıa, A., & Konig, P. (2018). Further advantages of data augmentation on convolutional neural networks. *International Conference on Artificial Neural Networks*, 95–103.

Jarrar, M. (2021). The Arabic Ontology – An Arabic Wordnet with Ontologically Clean Content. *Applied Ontology*, *16*(1), 1–26.

Joshi, R., Goel, P., & Joshi, R. (2020). Deep learning for hindi text classification: A comparison. *arXiv preprint arXiv:2001.10340*.

Kobayashi, S. (2018). Contextual augmentation: Data augmentation by words with paradigmatic relations. *In NAACL-HLT*.

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2020). A Survey on Text Classification: From Shallow to Deep Learning. *ACM Transactions on Intelligent Systems and Technology*, *37*(4). http://arxiv.org/abs/2008.00364

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.

Ramdhani, M. A., Maylawati, D. S. A., & Mantoro, T. (2020). Indonesian news classification using convolutional neural network. *Indonesian Journal of Electrical Engineering and Computer Science*, *19*(2), 1000-1009.

Porcu, S., Floris, A., & Atzori, L. (2020). Evaluation of Data Augmentation Techniques for Facial Expression Recognition Systems. *Electronics*, *9*(1892), 1–12. https://doi.org/10.3390/electronics9111892

Shoaib, M. R., Elshamy, M. R., Taha, T. E., El-fishawy, A. S., & El-Samie, F. E. A. (2022). Efficient deep learning models for brain tumor detection with segmentation and data augmentation techniques. *Concurrency Computat Pract Exper. 2022;E7031.*, *July*. https://doi.org/10.1002/cpe.7031

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(60), 1–48. https://doi.org/10.1186/s40537-019-0197-0

Ulrich, N., Goss, K.-U., & Ebert, A. (2021). Exploring the octanol–water partition coefficient dataset using deep learning techniques and data augmentation. *Communications Chemistry*, *4*(90), 1–10. https://doi.org/10.1038/s42004-021-00528-9

Uthayan, K. R., & Mala, G. S. A. (2015). Hybrid Ontology for Semantic Information Retrieval Model Using Keyword Matching Indexing System. *The Scientific World Journal*, *2015*(1).

Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* 6382–6388.

Xie, Z., Wang, S. I., Li, J., Levy, D., Nie, A., Jurafsky, D., & Ng, A. Y. (2017). *Data noising as smoothing in neural network language models.*

Yu, A. W., Dohan, D., Luong, M.-T., Zhao, R., Chen, K., Norouzi, M., & Le, Q. V. (2018). Qanet: Combining local convolution with global self-attention for reading comprehension. *CoRR, Abs/1804.09541*.