Natural Language Processing Approaches to Text Data Augmentation: A Computational Linguistic Analysis

DOI: https://doi.org/10.33806/ijaes.v25i1.682

Hoda Zaiton
Arab Academy for Science, Technology and Maritime Transport in Collaboration with Alexandria University, Egypt.
Sameh Alansary
Alexandria University, Egypt.

Received: 7.2.2024 Accepted: 11.6.2024 Published Online: 15.6.2024

Abstract: In the context of Natural Language Processing (NLP) tasks, problems such as insufficient or skewed data are frequently encountered. One practical solution to this problem is to generate additional textual data. Text Data Augmentation (TDA) refers to small changes made to accessible text at the character, word, or sentence level to generate synthetic data that is subsequently inserted into data loaders to train the model. By producing synthetic data, models can learn from a larger range of instances and, hence, enhance their resilience and generalization skills. Despite the fact that the entire NLP community has extensively studied many NLP DA approaches, recent research on the subject suggests that the relationship between the several DA techniques now in use is not entirely known in practice. Therefore, this study applies and extends the advances of TDA to encounter and cover varied tools on multiple settings or contexts. To carry out a thorough practical implementation of NLP DA approaches, comparing the way they perform and highlighting some of the significant similarities and differences in these various scenarios, this work depends on different tools of easy data augmentation and neural-based augmentation. This study suggests that some typical DA techniques might not be suitable in some circumstances or text environments. Specifically, according to the initial results, the context and word count of a text may have a significant impact on the quality of the synthetic data.

Keywords: easy data augmentation, natural language processing, neural augmentation, text data augmentation, text genre

1. Introduction

Data Augmentation (DA) is a technique that is generally used to create additional samples from existing ones, either by creating new synthetic data from already existing data or by expanding the dataset by making minor alterations or changes to the initial data. Accordingly, acts as a regularizer that reduces overfitting when training a machine learning model (Shorten, Khoshgoftaar and Furht 2021). DA is widely used for adding to image data, where it was proven to be effective (Szegedy et al. 2015). However, due to the complexity of the various languages and the range of NLP activities, DA is more challenging and less investigated in the subject of NLP. TDA is a method used for increasing the diversity of training

examples without explicitly collecting new data (Feng et al. 2021). Specifically, TDA is widely used in NLP applications that suffer from either a dearth of training data or the challenging in gathering of a large quantity of labelled data, including text classification (Xiang et al. 2021), named entity recognition (Sabty et al. 2021) and machine translation (Al-Taher 2019; Zhu et al. 2019).

Current NLP data augmentation methods function at multiple levels of granularity, including character, word and sentence or document. Textual augmentation at the character level (Belinkov and Bisk 2018; Karpukhin et al. 2019) introduces characters at arbitrary textual positions in a technique called Random Noise Injection (RNI). In this technique, noise is injected into the data through random character substitutions, random character removals, or random character flipping between two characters, all without considering the entire context. Keyboard augmentation and spelling (also known as spelling error insertion) augmentation (Coulombe 2018) are two sub-methods for the RNI technique. Whereas the keyboard augmentation technique replaces characters with adjacent characters to mimic typing errors or variances caused by keyboard depending on keyboard layouts, spelling augmentation generate variations of spelling errors that exhibit consistent patterns of misspellings, such as character substitutions, deletions and duplications.

Working at the word level for TDA (Kolomiyets, Bethard and Moens 2011; Wang and Yang 2015; Zhang, Zhao and LeCun 2015; Wei and Zou 2019) permits minor word-level alterations to be made to the text's content without altering its broader significance. In the light of that, similar adjustments can be demonstrated at two different levels: easy and advanced. Easy Data Augmentation (EDA) (Wei and Zou 2019) involves editing texts by adding, deleting, or substituting words. Advanced or complex techniques represent words using embeddings to insert and replace them according to the similarity they display with other words in the embedding space. Whilst word-level data augmentation focuses on introducing variations at the individual word level within sentences, sentence-level techniques, conversely, aim to create variations in the structure and content of sentences while conserving their original meaning. One popular method of text data augmentation at the sentence level is back-translation (Sennrich, Haddow and Birch 2015; Gangal et al. 2022), where a sentence is translated into another language and then back into the original language, resulting in a generation of diverse sentence-level variations. Another distinct approach is to reword statements while still retaining semantic equivalency by applying paraphrasing algorithms. Nonetheless, these approaches frequently result in creating multiple versions of a dataset, they are commonly unsuitable for sequential tagging applications that depend on the token-label association like semantic parsing applications. Furthermore, in conjunction with the aforesaid levels of augmentation (character, word, or sentence), TDA can be feasibly employed at the syntactic level. (Futrell, Mahowald and Gibson 2015; Gulordava et al. 2018; Sahin and Steedman 2018). The syntactic level of TDA involves manipulations in the grammar, syntax and structure of the original text to generate wide-ranging variations but preserving its underlying semantics. Thereby,

expanding the scope and quantity of the dataset for applications such as machine learning and NLP. One example of TDA at the syntactic level is Part-of-Speech (POS) tagging (Xiang et al. 2021; Kim, Won and Park 2022; Shen et al. 2022). POS tagging works by assigning parts of speech labels to words given in a text (Pandian and Geetha 2008). Those labels or "tags." can include negation and affirmation, pronoun and/or phrase replacements, or noun phrase generation in which other nouns, adjectives, prepositional phrases, or relative clauses are added to expand or modify a noun phrase.

Regardless of the text synthesis backbone or granularity level; easy or complicated, the goal of data augmentation is to generate logical, diversified and semantically coherent additional samples that effectively ensure both the quality and quantity of the training data. The quality of augmented textual data, including different variables such as the coherence, fluency, grammatical correctness and overall readability of augmented samples, holds paramount importance for the effectiveness of machine learning models and NLP systems and applications. Evaluating these language-related factors is essential for understanding the impact of augmentation techniques on the generated outputs as well as for ensuring the effectiveness of the augmented data. In order to address these issues, this study utilizes different text categories or genres to explore and investigate different types of augmentation methodologies that augment on the word (token) level only based on two approaches. The first approach works upon lexical substitution techniques based on the WordNet thesaurus (Miller 1995). From this approach we employ two augmenters; synonym augmenter (hereafter, SynonymAug) and antonym augmenter (hereafter, AntonymAug). The second approach operates on the neural augmentation (Shorten, Khoshgoftaar and Furht 2021) approach. Using this method, we apply two methods; a conventional approach based on GloVe (Pennington, Socher and Manning 2014) word embedding substitution (hereafter, GloVeAug); and transformer models augmentation based on BERT (Devlin, Chang, Lee and Toutanova 2019) Contextual Word Embeddings (CWE) (hereafter, BERT-CWEAug). Finally, upon performing that, the fundamental goal of this study is to evaluate the linguistic quality of the augmented texts to determine how the choice of TDA technique can affect the linguistic quality and coherence of the generated texts compared to the original ones. In other words, how well each TDA technique preserves the structural (i.e. data diversity and lexical coverage) as well as the functional (i.e. contextual relevancy) linguistic characteristics of the original text and whether they enhance or degrade its quality. Accordingly, this may help to reveal whether particular augmentation strategies are more appropriate or successful for particular text categories in light of their capacity to preserve and capture semantic and/or contextual similarity with the original text.

The remaining sections of the paper are arranged as follows: the theoretical framework supporting both DA and the linguistic properties of the chosen text types being assessed in this work is presented in *section 2*. The methodology is outlined in *section 3* and the evaluation standards are provided in *section 4*. The key findings and analytical decisions are presented in *section 5*. Before concluding, a discussion

derived from the assessment is presented in *section 6*. Finally, we conclude up in *section 7*, where we pinpoint certain intriguing avenues for further research.

2. Theoretical lens

2.1 Word-level augmentation

Word-level augmentation in NLP applications refers to the process of enhancing or modifying individual words within a text to generate new text. The essence of easy augmentation techniques of word-level augmentation involves word-by-word replacements through several modifications executed in a unified, cohesive manner to the source text in order to generate new synthetic texts (Pellicer, Ferreira and Costa 2023). Examples of these augmentation techniques are random swapping and random deletion (Wei and Zou 2019). Whilst swapping words is a sub-method in which specific words are provided and selectively exchanged in a text, random deletion works by applying a probabilistic parameter to the randomized deletion or removal of words from a phrase.

Another popular technique of word-level augmentation is lexical substitution-based lexicon entailing synonym and antonym substitution. Synonym substitution-based lexicon is a technique functions by replacing a random set of tokens from a text/document with another set of tokens with equal or equivalent meaning or synonyms using synonym databases or thesaurus such as PPDB (Pavlick et al. 2015), a paraphrase database that can be used to generate synthetic data and WordNet (Miller 1995), a lexical database for English. In NLP tasks such as text classification, sentiment analysis, or language generation, in particular, synonym substitution augmentation can be helpful since it exposes the model to a variety of word alternatives which enhances its performance and generalization skills. In the context of that, the work of Fadaee, Bisazza and Monz (2017) demonstrated that substitutions for typical synonyms produce superior performance in machine translation tasks, particularly for low-resource languages. They also noticed that, in machine translation tasks, substituting numerous words in a sentence outperforms changing just one. In the case of antonym substitution for TDA, random words from the sentence are replaced with their antonyms, using antonym database or thesaurus. In their work, Haralabopoulos, Torres, Anagnostopoulos and McAuley (2021) provide two text augmentation-based techniques, antonym and negation, that alter the classification of each augmented example. Furthermore, compared to permutation augmentation, a text augmentation technique that maintains all of the corpus' characteristics, including term frequency and class distribution, while simultaneously improving the classification results, they discovered that antonym and negation augmentations increase classification accuracy by a minimum of 0.35% (Haralabopoulos et al. 2021)

Unlike lexical replacements-based thesaurus or other random operations, additional word adaption algorithms functioning based on more intricate approaches or models to represent words through embeddings. Word-embeddings such as FastText embedding (Bojanowski, Grave, Joulin and Mikolov 2017) created by Facebook as an extension of Word2vec (Mikolov, Yih and Zweig 2013)

and GloVe (Pennington, Socher and Manning 2014) represent words with embeddings so that they can be exchanged out or added to a space based on how similar they are to other words. According to semantic similarity in the word embeddings, Wang and Yang (2015) presented a word embedding-based data augmentation method, such as Word2Vec or GloVe, that replaces words with their top-n comparable terms. In addition to capturing a greater variety of linguistic patterns, the models can be trained to handle variations in word choice, thus enabling effective replacements. The authors analyze this technique comparing it with training the same model without data augmentation and get statistically significant improved outcomes on classification models utilizing social media material. While Word2Vec, Glove and FastText exceed other basic or easy techniques at the word level, they are static models that remain constant regardless of context. Consequently, a novel word embeddings technique based on contextual data is presented. Contextual word embeddings provide contextualized word embeddings based on surrounding words in a sentence, allowing it to collect contextual information. CWE-based language models (LMs) for text augmentation can be implemented either to produce new words using traditional LMs such bi-directional LSTM-RNN (Kobayashi 2018), or by Transformer Models (TMs) such as BERT (Devlin et al. 2019), or GPT (Radford and Narasimhan 2018). TMs, by an expansion of neural networks, typically through techniques such as transfer learning (Pratt 1996) had undergone extensive training through the use of a pretext task known as Masked Language Modelling (MLM) to envisage masked words positions depending on the context. In the work of Wu et al. (2019), in order to get around some of the limitations in Kobayashi's (2018) LSTM-based LM, the authors propose conditional BERT. Given a labelled sentence, conditional BERT is used to predict new words that are compatible with the label of the given sentence after masking a few random words. This method yielded superior outcomes in all examined cases. In summary, contextual embeddings, as opposed to standard embedding models, generate context-dependent nature vectors for words (Sabty et al. 2021). This contextual model element aids in the modelling of complex and multidimensional words.

2.2 Language characteristics in different texts

Language appears in a variety of different categories or genres. Genres, in a broadly defined sense, are "conventional instances of organized text" Couture (1986: 80), that a writer produces for the demands of specific contexts (Johns 2008). A genre schema, according to this understanding, is defined by its function as well as its structure, content, lexis, grammar, layout, graphics and so on to include the possible contexts of occurrence (Bax 2011). To illustrate, genres such as business texts focus on topics such as market analysis, business strategies, industry trends and financial reports; thus, employ a formal tone of specialized terminology and jargon related to marketing, management and finance. In contrast, the entertainment genre, covers a wide range of topics such as movies, books, music, celebrities and popular culture. Therefore, it often uses a conversational and informal tone to connect with the

audience. Whilst, legal texts utilize precise and technical language to convey legal concepts and principles focusing on laws, regulations, court cases, legal procedures and interpretations (see Abu-Ssaydeh and Jarad 2016). Moreover, the complexity of written legal genre is displayed in the scope of the logic operations such as conditional, negation, conjunction and disjunction (Nazarenko and Wyner 2017). In the case of NLP augmentation, this variance in text can impact the performance of NLP models in argumentation tasks. To elaborate, models trained on one genre may not generalize well to other genres due to differences in vocabulary, tone, discourse patterns and the argumentative methods used to express a particular meaning. For instance, a model trained on legal texts may find it difficult to produce cogent arguments when it is applied to entertainment genres, which frequently have an informal tone. To tackle this difficulty, training NLP models on a variety of datasets spanning multiple genres is crucial.

In essence, to the best of our knowledge, no research has conducted a full analysis and comparison of NLP DA approaches in a transparent, replicable manner with extensive linguistic quality evaluation criteria. In the sections that follow, we start to address these research prospects by comparing several NLP DA techniques.

3. Methodology

3.1 Overall framework

Given the preceding literature review, this study seeks to point out some of the most patent research gaps found in the NLP DA literature. First, a dearth of comparative research studies suggesting a thorough examination of the linguistic quality of generated or augmented texts in comparison to the original text, assessing how well each augmentation tool performs for particular text kinds and examining how each text type responds to various augmentation techniques. Second, a comprehensive qualitative evaluation criterion of the artificially generated data in terms of sustaining its original substance; semantic similarity or contextual relevancy receive little to no attention. In contrast to what is common in image augmentation, NLP DA works rarely cover any topic other than model performance. Hence, considering the scarcity of works that take these variables into account and as previously stated, enhancing understanding of DA approaches in NLP, both in terms of qualitative assessment and practical implementation, constitutes an essential research topic. Finally, to achieve this goal, we implement in an open-source library, Python's nlpaug package (Ma 2019). Nlpaug is a rich and adaptable library that offers a wide range of augmentation techniques for text, audio and spectrogram data (Ma 2019). The entire operation is then completed using the following baseline method.

3.2 Baseline method

The baseline technique applied in this work is described as follows:

a. *Compiling or gathering a representative corpus* of texts or a dataset in the desired genre to serve as the basis for generating new texts. As such, we examined a corpus of six types of text: Business, entertainment, politics, sport,

technology and legal from two different datasets. The first five aforementioned categories are being obtained from BBC News dataset (Greene and Cunningham 2006), a public dataset originating from BBC News, consists of 2,225 documents, categorized into the five aforesaid groups. The sixth category, namely, legal is obtained from LexGLUE (Chalkidis et al. 2022) a Benchmark Dataset for Legal Language Understanding in English. LexGLUE is consisting of seven datasets: ECtHR (A), ECtHR (B), SCOTUS, EUR-LEX, LEDGAR, UNFAIR-ToS and CaseHOLD. Furthermore, for each genre or category, three separate samples with varying numbers of characters or tokens are used. Each sample is three times augmented or generated. The information about the compiled corpus is shown in Figure (1).

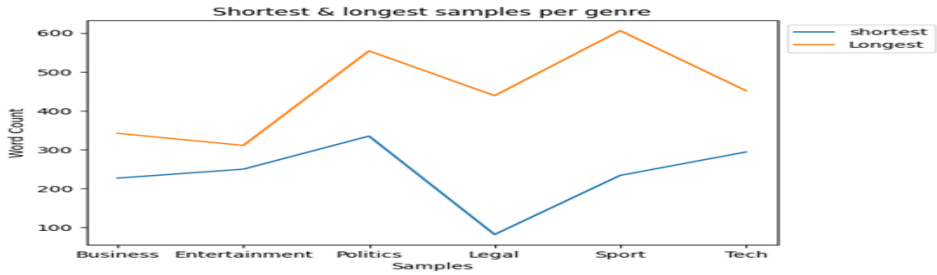

Figure 1. Word count representation for the compiled corpus; shortest and longest texts by word count.

Note: As seen in Figure (1), the shortest text of the business genre, for instance, is 227 by word count and the longest text is 342 by word count.

b. *Applying text data augmentation techniques* to bring changes to the original texts. In this work, the compiled corpus is generated and enhanced through the application of four tools from three distinct techniques. The first implemented technique is based word embedding substitution by GloVeAug. GloVeAug functions by substituting words in a sentence with similar meanings according to the semantic similarity of GloVe embeddings. The second technique operates by CWE approach utilizing BERT-CWEAug. Since, BERT model (Devlin et al. 2019) can predict the masked words based on the adjacent context, BERT-CWEAug works by capturing contextual information and producing embeddings that are relevant to the surrounding words. The third and fourth techniques put into practice; SynonymAug and AntonymAug, are predicated on WordNet thesaurus (Miller 1995). SynonymAug searches WordNet for synonyms or "synsets" for the words being augmented. These synsets are typically linked by lexical, conceptual and semantic relationships. On the other hand, AntonymAug replaces words with antonyms or opposites through utilizing the WordNet dataset.

c. *Iterate and refine* of the unsatisfactory outcome of the generated text. Within this study, a recurrence of the procedure by collecting two additional samples is deployed, resulting in a total of adjusting the augmentation approaches on three samples for each text category, each with three augmentation attempts. In

all circumstances, the study adheres to the third attempt for analysis and evaluation in order to retain the clarity and reliability of analysis.

d. *Evaluate the generated texts* using algorithmic and human evaluation standards to assess the fidelity and compactness of the freshly generated data. In profundity, the evaluation criteria performed are to be explained in the section that follows.

Finally, the expected step here, after evaluation of the generated texts, is to train the model on the target NLP task; however, in light of its goals and objectives, which are to ascertain how the methods used for TDA influences the linguistic coherence and quality of the enhanced texts compared to the original text, only phases covered prior to the model training step are used in this study.

4. Evaluation setup

4.1 Comparison and evaluation criteria

The assessment process was divided into two stages: machine and human evaluation. The selected applied comparison criteria are represented by the variables below and the evaluation technique is explained here.

1. *Semantic preservation*: refers to the ability each tool confronts to select words that embrace the similar or equivalent meanings between the original and augmented texts. To evaluate and assess the semantic similarity and preservation among texts, we employ cosine similarity matrix. Cosine similarity matrix is one of the most popular similarity metrics that is commonly used in NLP (Wang and Dong 2020). Cosine similarity measures the cosine value of the angle between two vectors, in this case, two documents, by calculating the distance relationship in the latent space. Accordingly, greater similarity between two vectors results in a higher score, suggesting better semantic preservation.
2. *Data diversity and lexical coverage*: refers to the ability of each tool to introduce diverse and rich semantic and syntactic variations in the augmented text data coverage. For this purpose, we employ one of the diversity metrics, measure of textual lexical diversity (MTLD) (Mccarthy 2005). MTLD measures the average length of sequential word segments in a text before reaching a pre-defined or additional threshold. Thus, MTLD captures variation within different sections of a text by considering both the number of types and tokens.
3. *Vocabulary expansion***:** refers to how effectively tools employed expands the vocabulary of the text data by introducing alternative word choices. Vocabulary expansion between the original and the synthetic texts is being counted and generated automatically with the augmented outputs.
4. *Contextual relevancy***:** refers to the ability of each tool to select words that are contextually relevant. Evaluation of contextual relevancy in augmented text data is carried out using a human-centered approach, as we believe that a human-centered method is more appropriate for quantifying this aspect. Furthermore, by accounting for the complex interplay of linguistic features and

contextual nuances inherent in the enhanced text data, this human-centric evaluation approach may ensure a refined analysis of contextual relevancy.

The subsequent section (section 5) delineates pragmatic deliberations pertaining to the implementation of text data augmentation techniques.

5. Experiment results

This section aims to give a thorough study of the performance across all the augmenters, taking into account both their influence on the quality of the created data and how well they enhanced the corpus. The results of the assessment were conducted in two stages and included both machine and human evaluations.

5.1 Machine evaluation

5.1.1 Semantic preservation

Scatterings presented in Figures (2, 3 and 4) show the cosine similarity scores and final accuracy for all data augmentation methods used on the developed corpus. Figure (2) presents cosine similarity measurements in sample 1, Figure (3) presents cosine similarity measurements in sample 2 and Figure (4) presents cosine similarity measurements in sample 3, respectively.

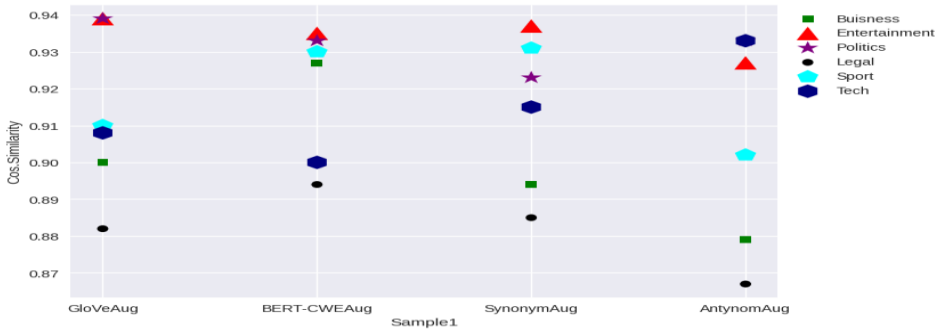

Figure 2. Cosine similarity in sample 1

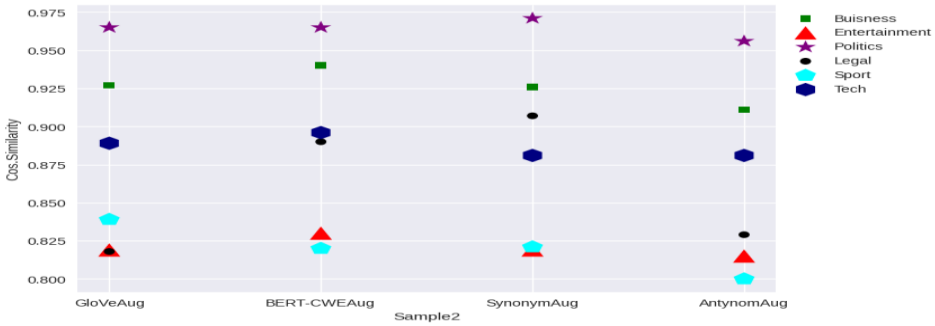

Figure 3. Cosine similarity in sample 2

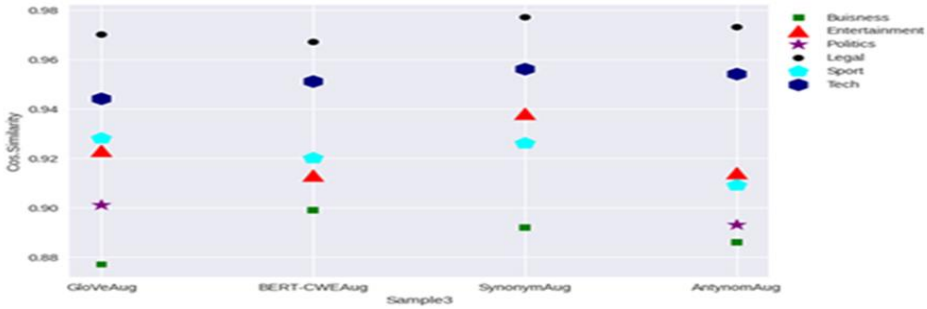

Figure 4. Cosine similarity in sample 3

As shown in Figures (2, 3 and 4), when compared to other text categories, the entertainment texts achieve the highest scores in sustaining semantic preservation in sample 1, however, when compared to other text categories in sample 2, the political text achieves the highest scores. Among text genres in sample 3, the legal genre obtains the highest scores for preserving semantic integrity. Based on the general observations in Figures (2, 3 and 4), Table (1) is provided.

Table 1. The depiction of cosine similarity among all samples

Text/Genre	Sample 1	Sample 2	Sample 3
Business	BERT-CWEAug	BERT-CWEAug	BERT-CWEAug
Entertainment	GloVeAug	BERT-CWEAug	SynonymAug
Politics	GloVeAug	SynonymAug	SynonymAug
Legal	BERT-CWEAug	SynonymAug	SynonymAug
Sport	GloVeAug	GloVeAug	GloVeAug
Tech	AntynomAug	GloVeAug	SynonymAug

From Table (1), three general findings can be drawn: **1)** *by means of values*, overall, six out of the eighteen texts provide the best performance when using SynonymAug, while the other six texts demonstrate the best performance when using GloVeAug. Five texts outperform with BERT-CWEAug and one text uses AntynomAug to prove the best performance: **2)** *by means of tool/augmenter*, GloVeAug surpasses other augmenters in sample 1, while the three augmenters of BERT-CWEAug, GloVeAug and SynonymAug compete and equate each other in sample 2. Additionally, SynonymAug achieves the best results in terms of semantic preservation between input and output texts in sample 3: **3)** *by means of responsiveness of a certain genre***,** business text uses BERT-CWEAug to demonstrate responsive performance, whereas the entertainment genre was responding to a variety of augmenters. Both politics and legal texts shows more response to SynonymAug, whereas sport text category responds significantly to GloVeAug. Finally, technology text depicts cosine similarity and sematic

preservation across all samples in a manner reminiscent of entertainment and responds to multiple augmenters.

5.1.2 Data diversity and lexical coverage

Figures (5, 6 and 7) show the final accuracy and MTLD values across all data augmentation techniques on the compiled corpus. Figure (5) grants MTLD in sample 1, Figure (6) presents MTLD in sample 2 and Figure (7) shows MTLD in sample 3, respectively.

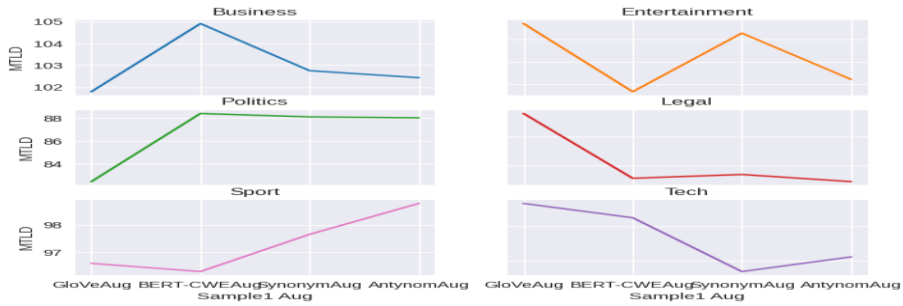

Figure 5. MLTD in sample 1

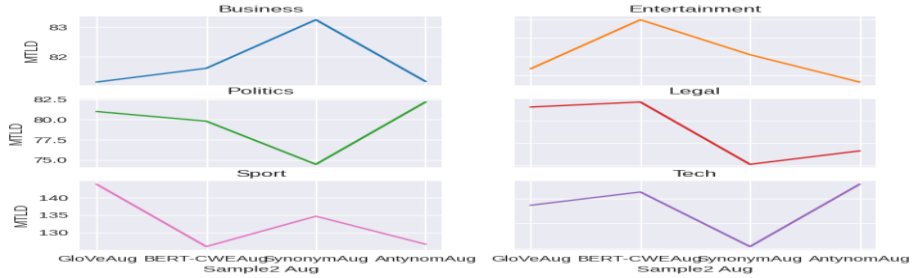

Figure 6. MLTD in sample 2

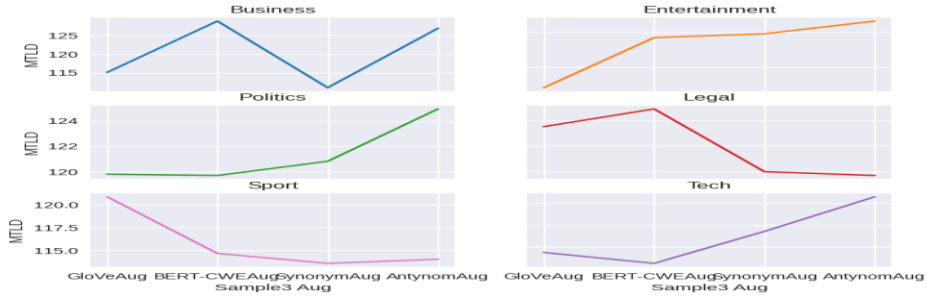

Figure 7. MLTD in sample 3

As seen in Figures (5, 6 and 7), taking the business genre as an example across the three samples, we regard that in sample 1, among all augmenters, BERT-CWEAug obtains the highest level in the coverage of lexical and data variety. In sample 2, SynonymAug outperforms other tools and once again, in sample 3, BERT-CWEAug beats other tools in covering a variety of lexical data in business text. Further, Table (2) provides a thorough illustration of the data variety and

lexical coverage performance obtained across the texts by all augmentation techniques.

Table 2. The depiction of data diversity and lexical coverage among all samples

Text/Genre	Sample 1	Sample 2	Sample 3
Business	BERT-CWEAug	SynonymAug	BERT-CWEAug
Entertainment	GloVeAug	BERT-CWEAug	AntynomAug
Politics	BERT-CWEAug	AntynomAug	AntynomAug
Legal	GloVeAug	BERT-CWEAug	BERT-CWEAug
Sport	AntynomAug	GloVeAug	GloVeAug
Tech	GloVeAug	AntynomAug	AntynomAug

Based on the information contained in Table (2), we can draw three general conclusions on the evaluation of the MTLD among texts: **1)** *by means of values*, six texts out of eighteen show the best performance utilizing BERT-CWEAug and similarly, six texts show the highest performance when using AntynomAug. While five texts outperform with GloVeAug and one text uses SynonymAug to attain lexical coverage and data variety: **2)** *by means of tool/augmenter*, GloVeAug outpaces other augmenters in sample 1, whilst BERT-CWEAug and AntynomAug compete and equate each other in sample 2 and AntynomAug exceeds other tools in sample 3 in introducing diverse lexical data: **3)** *by means of responsiveness of a certain genre***,** overall, business text responds to BERT-CWEAug, whereas entertainment genre responds to a variety of augmenters. Politics texts respond more to AntynomAug and legal texts respond to BERT-CWEAug. Sport text category shows great response to GloVeAug, while technology text resembles politics and shows response almost to AntynomAug in the depiction of data diversity and lexical coverage among all samples.

5.1.3 Vocabulary expansion

Plottings in Figures (8, 9 and 10) depict the vocabulary expansion measurement and values obtained across all techniques applied to the developed corpus. Figure (8) grants vocabulary expansion measurement in sample 1, Figure (9) presents vocabulary expansion rate in sample 2 and Figure (10) shows vocabulary expansion scores in sample 3, respectively.

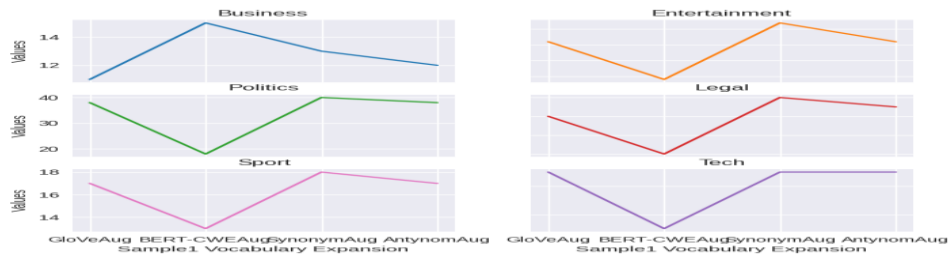

Figure 8. Vocabulary expansion in sample 1

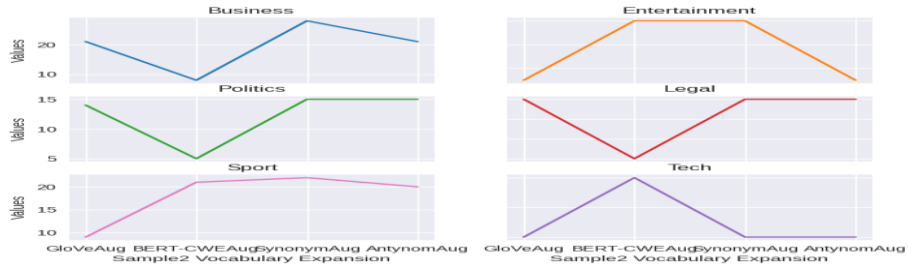

Figure 9. Vocabulary expansion in sample 2

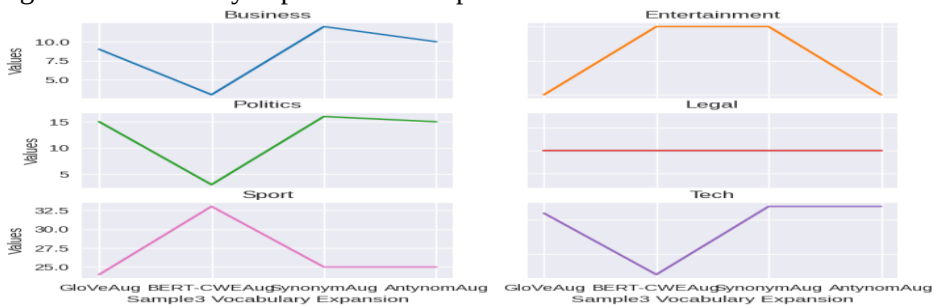

Figure 10. Vocabulary expansion in sample 3

As shown in Figures (8, 9 and 10), three general conclusions concerning the three samples of the vocabulary expansion measurement among the original and supplemented texts can be drawn: **1)** *by means of values*, of the eighteen texts, sixteen among them leverage SynonymAug to get the best results, either by outpacing other augmenters or by equating values with one or more augmenter. Five texts show the best performance by utilizing AntynomAug and six by BERT-CWEAug. However, when evaluating vocabulary expansion with GloVeAug, three texts do the best: **2)** *by means of tool/augmenter*, SynonymAug outperforms its counterparts across all three samples by significantly enhancing the lexical diversity within augmented texts as compared to their respective original versions through the introduction of alternative lexical selections. BERT-CWEAug and AntynomAug occupy the second position, while GloVeAug ranks third in terms of efficacy: **3)** *by means of responsiveness of a certain genre***,** the two aforementioned observations suggest that SynonymAug exhibits greater responses from all text categories than other augmenters.

5.2 Human evaluation

The primary objective of deploying human evaluation methodology in the present study revolves around assessing specific performance criteria pertaining to contextual relevance of textual content. For visual enrichment upon the contextual relevance of textual content, Tables (3, 4, 5 and 6) present an excerpt from a political text, accompanied by its corresponding augmented outputs across all augmenters applied. These tables include word counts for both the original text and

the generated versions. Notably, words rendered in *italics* within the augmented extracts denote word substitutions generated.

5.2.1 Contextual relevancy

Upon careful analysis of the generated outputs using each respective tool, this investigation has revealed certain key points: First, although GloVeAug leverages GloVe embeddings to introduce semantically similar and diverse word choices or synonyms into the text data, in most instances, GloVeAug did not take into account the contextual meaning of words within the surrounding text. Consequently, irrelevant contextual modifications are generated. This might be the case because GloVeAug does not consider the contextual meaning of words inside the surrounding text; instead, it only considers word similarities in the embedding space. Table (3) presents the original excerpt accompanied by its corresponding augmented output by GloVeAug.

Table 3. The original and the generated augmented extract by GloVeAug

Original: 394	Augmented: 408
Tory leader urges Harry apology Prince Harry should personally make clear how sorry he is about wearing a Nazi uniform to a friend's fancy dress party, says Tory leader Michael Howard. Mr Howard, whose grandmother died in Auschwitz, said many people would be offended by the prince's actions Clarence House has issued a statement saying the prince has apologised and realised it had been a poor costume. Number 10 said an error was made but now Harry had apologised the matter should be left to the palace That was a message repeated by Home Secretary Charles Clarke who said the matter should now be left to lie. But Lib Dem leader Charles Kennedy said Harry was in a privileged position and said he should apologise in person.	Tory leader urges Harry apology Prince Harry should personally make clear how sorry he is about wearing a Nazi uniform to a friend's fancy dress party, says Tory leader Michael Howard. Mr Howard, whose grandmother died in Auschwitz, said many people *must* be offended by the prince's actions Clarence House has issued a statement saying the prince has apologised and realised it had been a poor costume. Number 10 said *a* error was made but now Harry had apologised the matter should be left to the *ducal***.** That was a message repeated by Home Secretary Charles Clarke who said the matter should now be *on* to lie. But Lib Dem *accused* Charles Kennedy said Harry was in a privileged position and said he should apologise in person.

Second, nonetheless, BERT-CWEAug, in most cases, generates limited patterns of augmented data; yet, those limited patterns produce natural-sounding augmented sentences that closely resembles the original text, resulting in a generation of relevant contextual information. In light of this, we find that BERT-CWEAug augments entire sentences rather than individual words or phrases,

thereby preserving the original text's coherence and contextual meaning. Table (4) displays the original excerpt accompanied by its corresponding augmented output by BERT-CWEAug.

Table 4. The original and the generated augmented extract by BERT-CWEAug

Original: 394	Augmented: 399
Tory leader urges Harry apology Prince Harry should personally make clear how sorry he is about wearing a Nazi uniform to a friend's fancy dress party, says Tory leader Michael Howard. Mr Howard, whose grandmother died in Auschwitz, said many people would be offended by the prince's actions Clarence House has issued a statement saying the prince has apologised and realised it had been a poor costume. Number 10 said an error was made but now Harry had apologised the matter should be left to the palace That was a message repeated by Home Secretary Charles Clarke who said the matter should now be left to lie. But Lib Dem leader Charles Kennedy said Harry was in a privileged position and said he should apologise in person.	Tory leader urges Harry apology Prince Harry should personally make clear how sorry he is about wearing a Nazi uniform *into* a friend's fancy dress party, says Tory leader Michael Howard. Mr Howard, whose grandmother died in Auschwitz, said many people would be offended by the prince's actions Clarence House has issued a statement saying the prince has apologised and realised it had been a poor costume. Number 10 said an error was made but now Harry had apologised the matter should be left to the palace. That *included* a message repeated by Home Secretary Charles Clarke who said the matter should now be left to lie. But Lib Dem leader Charles Kennedy said Harry was in a privileged position and said he should *apologize* in person

Third, although SynonymAug leverages thesauri to select synonyms that are contextually more relevant, in few cases, SynonymAug occasionally provides inappropriate synonym replacements, resulting in irrelevant context and the creation of unnatural-sounding sentences. Furthermore, we observe that certain synonyms have numerous meanings, making some cases contextually unclear. For example, SynonymAug replaces the word "*party*" in Table (5) with "*company*", which shows a futility in the original context. One way to explain this is that the word "*party*" has numerous connotations and SynonymAug failed to uncover contextually relevant synonyms for the word "*party*". Table (5) introduces the original excerpt accompanied by its corresponding augmented output by SynonymAug.

Table 5. The original and the generated augmented extract by SynonymAug

Original: 394	Augmented: 409
Tory leader urges Harry apology Prince Harry should personally make clear how sorry he is about wearing a Nazi uniform to a friend's fancy dress party, says Tory leader Michael Howard. Mr Howard, whose grandmother died in Auschwitz, said many people would be offended by the prince's actions Clarence House has issued a statement saying the prince has apologised and realised it had been a poor costume. Number 10 said an error was made but now Harry had apologised the matter should be left to the palace. That was a message repeated by Home Secretary Charles Clarke who said the matter should now be left to lie. But Lib Dem leader Charles Kennedy said Harry was in a privileged position and said he should apologise in person.	Tory *drawing card* urges Harry apology Prince Harry should personally make clear how sorry he is about wearing a Nazi uniform to a friend's fancy dress *company*, says Tory leader Michael Howard. Mr Howard, whose grandmother died in Auschwitz, said many people would be offended by the prince's actions Clarence House has issued a statement saying the prince has apologised and realised it had been a poor costume. Number 10 said an error was made but now Harry had apologised the matter should be left to the palace. That be a message repeated by Home Secretary Charles Clarke who said the matter should now be left to lie. But Lib Dem leader Charles Kennedy said Harry was in a privileged position and said he should apologise in person.

Fourth, AntynomAug allows for the exploration of opposite meanings or sentiments within the text; yet, replacing words with antonyms, in most cases, may not always preserve the overall context or meaning of the original text. In light of this, an exploration of opposite meanings or sentiments within the text can be valuable for certain tasks such as sentiment analysis and text classification where understanding contrasting perspectives is important. Therefore, we may not depend on AntynomAug method to preserve the contextual relevancy except for some cases and tasks such as those previously mentioned. Table (6) grants the original excerpt accompanied by its corresponding augmented output by AntonymAug.

Table 6. The original and the generated augmented extract by AntonymAug

Original: 394	Augmented: 409
Tory leader urges Harry apology Prince Harry should personally make clear how sorry he is about wearing a Nazi uniform to a friend's fancy dress party, says Tory leader Michael Howard. Mr Howard, whose grandmother died in Auschwitz, said many people would be offended by the prince's actions Clarence House has issued a statement saying the prince has apologised and realised it had been a poor costume. Number 10 said an error was made but now Harry had apologised the matter should be left to the palace. That was a message repeated by Home Secretary Charles Clarke who said the matter should now be left to lie. But Lib Dem leader Charles Kennedy said Harry was in a privileged position and said he should apologise in person.	Tory leader urges Harry apology Prince Harry should personally *break* clear how sorry he is about wearing a Nazi uniform to a friend's fancy dress party, says Tory leader Michael Howard. Mr Howard, whose grandmother died in Auschwitz, said *few* people would be offended by the prince's actions Clarence House has issued a statement saying the prince has apologised and realised it had *differ* a poor costume. Number 10 said an error was made but now Harry had apologised the matter should be left to the palace. That was a message repeated by Home Secretary Charles Clarke who said the matter should now be left to lie. But Lib Dem leader Charles Kennedy said Harry *differ* in a privileged position and said he should apologise in person.

To summarize, based on an exhaustive examination of the generated outputs produced by each tool, we observe that BERT-CWEAug surpasses and exceeds other augmenters in capturing more accurate meaning and context of words. Similarly, BERT-CWEAug exceeds other tools in generating expressive sentences. Through this, the coherence and contextual significance of original text are preserved.

6. Discussion

In this study, our concern is the augmentation of textual data, specifically, working at the word-level augmentation. Working at the word-level augmentation introduces variations in the textual content. While some of these variations preserve the syntax or semantics of the original sentence, the majority does not. Moreover, certain augmenters provide alternative word decisions, thereby introducing various variations of word alternatives. These alternative word choices primarily encompass adjectives, nouns, verbs and other linguistic categories. Overall observations upon the conducted analysis reveals several issues: First, out of all the augmenters used within this work, the best augmenter in keeping the semantic content and tone between original and augmented texts is SynonymAug and GloVeAug comes next in performance. Third position goes to BERT-CWEAug, while AntynomAug performs the lowest, achieving the fourth place. Second,

enhancing the lexical diversity in augmented texts is mostly achieved by the application of BERT-CWEAug and AntynomAug tools. Specifically, the AntynomAug tool exhibits superior performance in this regard, with BERT-CWEAug ranking second, while GloVeAug follows closely behind. Conversely, SynonymAug demonstrates comparatively less efficacy in facilitating lexical diversity and coverage measurements. In light of that, BERT-CWEAug predominantly depend on the substitution and provision of alternative verbs to enrich the text. Among the myriad of verbs substituted by BERT-CWEAug, verbs "*save*" to "*revive*", "*distributed*" to" *detonated*", "*told*" to "*informed*" and "*employ*" to "*hire*". Similarly, AntynomAug predominantly utilizes the strategy of verb substitution to introduce a numerous of variations. Among these verbs replaced by AntynomAug, verbs "*deny*" to "*allow*", "*promote*" to "*demote", "employ"* to *"fire" and* "*stay*" to "*depart*". Even though, AntynomAug mostly enhances the variations of lexical opposites, ensuring a more accurate alignment with their respective opposite meanings compared to their original counterparts, in most cases, it fails to correctly align the original verb tense with an appropriate substitution, leading to frequent grammatical errors or inconsistencies in the augmented text. In other words, a substantial prevalence of inaccurate verb tense matching is discernible. Among the substituted verbs that exhibit a lack of concordance in terms of tense, verb "*shipped*" to "*disembark*", "*rejects*" to "*admit*", "*drawn*" to "*deposit*", "*includes*" to "*exclude*", "*disobeys*" to "*obey*", "*endured*" to "*enjoy*" and "*won*" to "*lose*". Furthermore, in instances where AntynomAug fails to identify suitable opposite alternatives for words within the original text, the term "*differ*" is inserted as a replacement.

Third, by offering substitute word possibilities, SynonymAug excels other tools in terms of extending and expanding the vocabulary of the text. These substitute word alternatives are primarily introduced by adjectives. Among these adjectives substituted by SynonymAug are "*the past*" to "*the preceding*", "*new*" to "*fresh*", "*same*" to "*like*", "*small*" to "*little*" and "*financial*" to "*fiscal*". When comparing the rates at which vocabulary is expanded in the generated text, BERT-CWEAug and AntynomAug come in second place and GloVeAug takes the third place. Furthermore, a notable positive correlation is observed between the length, or word count, of a given text and the extent of vocabulary expansion in augmented versions of that text. Put simply, as the word count of the original text increases, the rate of expansion in the generated text also increases and vice versa. It is also worth noting that all augmenters employed in this study introduce additional spaces around commas, apostrophes and full stops, thereby increasing the overall length of the text segments. Fourth, when using BERT for data augmentation, it uses its understanding of language and semantics to construct contextual and meaningful sentences that closely resemble the original text, yielding contextually relevant sentences. Additionally, in maintaining relevant context, GloVeAug achieves the second place, followed by SynonymAug. In contrast, by presenting opposing selections, AntynomAug earns the lowest and final place in the maintaining of context.

Finally, in regard to the responsiveness of certain type of genre to a certain augmenter and based on both the deployed machine evaluation and the human review, this study observes that *business* text shows responsive performance mainly with BERT-CWEAug, while *entertainment* genre was responsive to multiple augmenters, in particular, GloVeAug and BERT-CWEAug. Texts on *politics* and *legal* respond more strongly to SynonymAug, though the *sport* text category responds well to GloVeAug, but the *technology* text category responds roughly equally to GloVeAug and AntynomAug. Indicating that each aforementioned tool or model might be trained and generalized more on specific genre(s) or domain. In this context, BERT-CWEAug, for example, might be generalized more on the *business* genre than other genres. Additionally, this may indicate that a specific tool may not be able to produce well-reasoned arguments when used with a particular genre of text. This explains why working with a certain sort of text might have an impact on how well NLP models or tools perform.

7. Conclusion

This work has explored a wide range of text augmentation methodologies that augment on the word (token) level only. Thus, we apply a diverse set of text genres to determine how the choice of text data augmentation technique can affect the linguistic quality and coherence of the augmented texts compared to the original texts. Furthermore, the conclusions we draw about a text data augmentation method from how well it works with a certain text genre. Upon conducting a comparison examination of specific language assessment factors and criteria, such as semantic similarity preservation, data diversity and lexical coverage, vocabulary expansion and contextual relevancy, this study concludes that, in all the cases, when it comes to text data augmentation, the choice of technique plays a vital role in determining the quality and relevance of the augmented data. Therefore, it is crucial to first define the NLP task before implementing text data augmentation, as some techniques might be more suitable for certain tasks than others. For example, under a circumstance of working on tasks related to machine translation, it is unsuitable to utilize AntynomAug. Furthermore, the evidence and analysis presented strongly support the notion that text data augmentation techniques can potentially reveal characteristics of certain genres or texts. Concerning that, through the application of augmentation techniques in a particular genre, scholars are able to discern patterns and variances within the augmented texts, hence, providing insights into the linguistic and stylistic characteristics of the original genre. Overall, for training NLP models on a broader range of data and improving their ability to handle different language variations, it is recommended by this study to start with synonym augmentation technique due to its ability to expand the vocabulary of the text data by introducing alternative word choices. Then, a customized enhancement strategy depending on the task can be applied. In conclusion, the offered data and analysis firmly bolster the idea that the choice of text data augmentation technique can have a significant impact on the linguistic quality and coherence of the augmented texts

compared to the original text. Different augmentation techniques employ various strategies to modify or generate new text and their effectiveness in preserving quality and coherence can vary. In future work, we seek to investigate which text genres present the biggest challenges for text data augmentation methods to advance in. In parallel, we aim to examine the efficacy of generative AI to broaden the benchmark of approaches and broaden our investigation to include other NLP tasks.

Hoda Zaiton- Corresponding Author
MA Candidate- Applied Linguistics
College of Language and Communication (CLC)
The Arab Academy for Science, Technology and Maritime Transport (AASTMT), in collaboration with the Institute of Applied Linguistics and Translation, Faculty of Arts, Alexandria University, Egypt.
ORCID Number: 0009-0008-7890-6220
E-mail: hoda.zaitoon@alexu.edu.eg

Sameh Alansary
Professor of Computational Linguistics
Phonetics and Phonology Department
Faculty of Arts, Alexandria University, Egypt.
ORCID Number: 0009-0007-2950-9555
E-mail: S.alansary@alexu.edu.eg

References

Abu-Ssaydeh, Abdul-Fattah and Najib Jarad. (2016). 'Complex sentences in English legislative texts: Patterns and translation strategies'. *International Journal of Arabic-English Studies* (*IJAES*)*,* 16(1): 111-128.

Al-Taher, Mohammad Anwar. (2019). 'Google translate's rendition of verb-subject structures in Arabic news reports'. *International Journal of Arabic-English Studies* (*IJAES*), 19(1):195-208. https://doi.org/10.33806/ijaes.19.1.11.

Bax, Stephen. (ed.). (2011). *Discourse and Genre*. London: Macmillan Education UK.

Belinkov, Yonatan and Yonatan Bisk. (2018). 'Synthetic and natural noise both break neural machine translation'. In *6th International Conference on Learning Representations, ICLR 2018, Conference Track Proceedings*, 1–13.

Bojanowski, Piotr, Edouard Grave, Armand Joulin and Tomas Mikolov. (2017). 'Enriching word vectors with subword information'. *Transactions of the Association for Computational Linguistics,* 5: 135–46. https://doi.org/10.1162/tacl_a_00051.

Chalkidis, Ilias, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz and Nikolaos Aletras. (2022). 'LexGLUE: A benchmark dataset for legal language understanding in English'. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume1: Long Papers)*, 4310–4330. https://doi.org/10.18653/v1/2022.acl-long.297

Coulombe, Claude. (2018). 'Text data augmentation made simple by leveraging NLP Cloud APIs'. Available at: http://arxiv.org/abs/1812.04718

Couture, Barbara. (1986). 'Effective ideation in written text: A functional approach to clarity and exigence'. *Faculty Publications--Department of English*. 67: 69-92.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. (2019). 'BERT: Pre-Training of deep bidirectional transformers for language understanding'. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Fadaee, Marzieh, Arianna Bisazza and Christof Monz. (2017). 'Data augmentation for low-resource neural machine translation'. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Volume 2: Short Papers*, 567–573. https://doi.org/10.18653/v1/P17-2090

Feng, Steven Y., Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura and Eduard Hovy. (2021). 'A survey of data augmentation approaches for NLP'. In *Findings of the Association for*

Computational Linguistics: ACL/IJCNLP, 968–988. https://doi.org/10.18653/v1/2021.findings-acl.84/

Futrell, Richard, Kyle Mahowald, and Edward Gibson. (2015). 'Quantifying word order freedom in dependency corpora'. In *Proceedings of the Third International Conference on Dependency Linguistics*, 91–100.

Gangal, Varun, Steven Y. Feng, Malihe Alikhani, Teruko Mitamura and Eduard Hovy. (2022). 'NAREOR: The narrative reordering problem'. *Proceedings of the 36th AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 36 (10): 10645–53. https://doi.org/10.1609/aaai.v36i10.21309.

Greene, Derek and Pádraig Cunningham. (2006). 'Practical solutions to the problem of diagonal dominance in Kernel document clustering'. In *Proceedings of the 23rd International Conference on Machine Learning-ICML'06*,337-384. https://dl.acm.org/doi/10.1145/1143844.1143892

Gulordava, Kristina, Piotr Bojanowski, Edouard Grave, Tal Linzen and Marco Baroni. (2018). 'Colorless green recurrent networks dream hierarchically'. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, Volume 1 (Long Papers)*, 1195–1205. https://doi.org/10.18653/v1/N18-1108

Haralabopoulos, Giannis, Mercedes Torres Torres, Ioannis Anagnostopoulos and Derek McAuley. (2021). 'Text data augmentations: Permutation, antonyms and negation'. *Expert Systems with Applications*, 177 : 114769. https://doi.org/10.1016/j.eswa.2021.114769

Johns, Ann M. (2008). 'Genre awareness for the novice academic student: An ongoing quest'. *Language Teaching*, 41 (2): 237–52. https://doi.org/10.1017/s0261444807004892

Karpukhin, Vladimir, Omer Levy, Jacob Eisenstein and Marjan Ghazvininejad. (2019). 'Training on synthetic noise improves robustness to natural noise in machine translation'. In *Proceedings of the 5th Workshop on Noisy User-generated Text, W-NUT@EMNLP 2019*, 42–47. https://doi.org/10.18653/v1/D19 -5506

Kim, Hyeon Soo, Hyejin Won, and Kyung Ho Park. (2022). 'PMixUp: Simultaneous utilization of part-of-speech replacement and feature space interpolation for text data augmentation'. https://openreview.net/forum?id=O4fNuE8F51T

Kobayashi, Sosuke. (2018). 'Contextual augmentation: Data augmentation by words with paradigmatic relations'. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 452–457. https://doi.org/10 .18653/v1/N18-2072

Kolomiyets, Oleksandr, Steven Bethard, and Marie-Francine Moens. (2011). 'Model-Portability experiments for textual temporal analysis'.

In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2: 271–276.

Ma, Edward. (2019). *Nlpaug: Data Augmentation for NLP*. Available at: https://github.com/makcedward/nlpaug

Mccarthy, Philip M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity. PhD Dissertation, The University of Memphis.

Mikolov, Tomáš, Wen-tau Yih, and Geoffrey Zweig. (2013). 'Linguistic regularities in continuous space word representations'. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 7: 46–51.

Miller, George A. (1995). 'WordNet: A lexical database for English'. *Communications of the ACM*, 38 (11): 39–41. https://doi.org/10.1145/219717.219748

Nazarenko, Adeline, and Adam Wyner. (2017). 'Legal NLP Introduction'. *Traitement automatique des langues*, 58(2): 7-19.

Lakshmana Pandian, S., and T. V. Geetha. (2008). 'Morpheme based language model for Tamil part-of-speech tagging'. *Polibits*, 38: 19–25.

Pavlick, Ellie, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme and Chris Callison-Burch. (2015). 'PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings and style classification'. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 425-430.

Pellicer, Lucas Francisco Amaral Orosco, Taynan Maier Ferreira and Anna Helena Reali Costa. (2023). 'Data augmentation techniques in natural language processing'. *Applied Soft Computing*, 132: 109803. https://doi.org/10.1016/j.asoc.2022.109803 .

Pennington, Jeffrey, Richard Socher and Christopher Manning. (2014). 'Glove: Global vectors for word representation'. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543. https://doi.org/10 .3115/v1/D14-1162

Pratt, Lorien. (1996). 'Special issue: Reuse of neural networks through transfer'. *Connection Science, (Print)*, 8(2).

Radford, Alec and Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. (2018). 'Improving language understanding by generative pre-training'. Available at https://api.semanticscholar.org/CorpusID:49313245.

Sabty, Caroline, Islam Omar, Fady Wasfalla, Mohamed Islam and Slim Abdennadher. (2021). 'Data augmentation techniques on Arabic data for named entity recognition'. *Procedia Computer Science*, 189: 292–99. https://doi.org/10.1016/j.procs.2021.05.092.

Şahin, Gözde Gül, and Mark Steedman. (2018). 'Data augmentation via dependency tree morphing for low resource languages'. In *Conference on*

Empirical Methods in Natural Language Processing, 5004–5009. https://doi.org/10.18653/v1/D18-1545

Sennrich, Rico, Barry Haddow and Alexandra Birch. (2015). 'Improving neural machine translation models with monolingual data'. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 86–96. https://doi.org/10.18653/v1/P16-1009

Shen, Yutong, Jiahuan Li, Shujian Huang, Yi Zhou, Xiaopeng Xie, and Qinxin Zhao. (2022). 'Data augmentation for low-resource word segmentation and pos tagging of ancient Chinese texts'. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, 169–173.

Shorten, Connor, Taghi M. Khoshgoftaar and Borko Furht. (2021). 'Text data augmentation for deep learning'. *Journal of Big Data*, 8 (1): 101. https://doi.org/10.1186/s40537-021-00492-0 .

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke and Andrew Rabinovich. (2015). 'Going deeper with convolutions'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1-9.

Wang, Jiapeng and Yihong Dong. (2020). 'Measurement of text similarity: A survey'. *Information*, 11 (9): 421. https://doi.org/10.3390/info11090421.

Wei, Jason and Kai Zou. (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.

Wang, William and Diyi Yang. (2015). 'That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets'. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2557–2563.http://dx.doi.org/10.18653/v1/D15-1306

Wu, Xing, Shangwen Lv, Liangjun Zang, Jizhong Han and Songlin Hu. (2019). 'Conditional BERT contextual augmentation'. In *Computational Science - ICCS 2019 - 19th International Conference, Proceedings, Part IV*, 84–95. http://dx.doi.org/10.1007/978-3-030-22747- 0_7.

Xiang, Rong, Emmanuele Chersoni, Qin Lu, Chu-Ren Huang, Wenjie Li and Yunfei Long. (2021). 'Lexical data augmentation for sentiment analysis'. *Journal of the Association for Information Science and Technology*, 72 (11): 1432–47. https://doi.org/10.1002/asi.24493.

Zhang, Xiang, Junbo Zhao and Yann LeCun. (2015). 'Character-level convolutional networks for text classification'. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*, 649–657.

Zhu, Jinhua, Fei Gao, Lijun Wu, Yingce Xia, Tao Qin, Wengang Zhou, Xueqi Cheng and Tie-Yan Liu. (2019). 'Soft contextual data augmentation for neural machine translation'. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5539–5544. http://dx.doi.org/10.18653/v1/P19-1555