



ISLAMIC UNIVERSITY OF TECHNOLOGY

**A Machine Learning approach to Data
Augmentation with Semantic Similarity on a
Low-Resource Language**

By

Shah Jawad Islam	180041223
Mohammad Abrar Chowdhury	180041235
Taufiqul Alam	180041236

Supervisors

Dr. Hasan Mahmud
Associate Professor

Nafisa Sadaf
Lecturer

Dr. Md. Kamrul Hasan
Professor

Systems and Software Lab (SSL)
Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT)
A subsidiary organ of the Organization of Islamic Cooperation (OIC)

*A thesis submitted in partial fulfilment of the requirements
for the degree of B.Sc. in Computer Science and Engineering*

Academic Year: 2021-2022

May 2023

Abstract

The augmentation of data in low-resource languages gained significant importance recently, primarily because of scarcity of datasets or the presence of highly unbalanced datasets. In the case of the Bengali language, the detection of fake news has turned up as a relevant problem, particularly in light of the surge in false information related to Covid-19 and the pandemic [1]. However, there has been a lack of adequately balanced data sets specifically designed for training Machine Learning (ML) and Deep Learning (DL) models in the detection of fake news in Bengali. Furthermore, previous attempts at augmenting fake news texts have yielded satisfactory results in lexical analysis but unsatisfactory results in terms of semantic relevance. To address these challenges, we propose a framework that involves the use of Text Augmentation techniques with the assistance of the Bangla Text-to-Text Transfer Transformer (T5) model. This framework aims to balance an unbalanced Bengali fake news dataset, while ensuring that the augmented text retains semantic similarity and structural accuracy. By employing this approach, we seek to strengthen the effectiveness and reliability of fake news detection models in the Bengali language.

Keyword - Text Augmentation; Balanced Dataset; Lexical Analysis; Semantic Relevance; Bangla T5

Declaration of Authorship

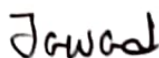
This is to declare that the work that has been presented in this thesis is the result of the experiments and analysis done out by Taufiqul Alam, Shah Jawad Islam and Md. Abrar Chowdhury under the supervision of Dr. Hasan Mahmud, Associate Professor and Nafisa Sadaf, Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka. It is also stated that this thesis or any portion of it has not been previously submitted for any degree or diploma. The work acknowledges and references information obtained from the published and unpublished works of others.

Authors:



Taufiqul Alam

Student ID - 180041236



Shah Jawad Islam

Student ID - 180041223



Mohammad Abrar Chowdhury

Student ID - 180041235

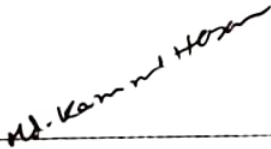
May, 2023

Supervisors:

Dr. Hasan Mahmud
Associate Professor
Systems and Software Lab (SSL)
Department of Computer Science and Engineering
Islamic University of Technology (IUT)



Nafisa Sadaq
Lecturer
Systems and Software Lab (SSL)
Department of Computer Science and Engineering
Islamic University of Technology (IUT)



Dr. Md. Kamrul Hasan
Professor
Systems and Software Lab (SSL)
Department of Computer Science and Engineering
Islamic University of Technology (IUT)

Acknowledgements

We would like to express our heartfelt thanks to Allah Subhanu Wata'ala for giving us the strength to finish this study and being with us.

We are very thankful to **Dr Hasan Mahmud**, Associate Professor, Department of Computer Science & Engineering, IUT for guiding and mentoring us throughout our research journey. His motivation, suggestions, and profound insights have played a pivotal role in the success of this study. Without his unwavering support and expert guidance, this research would not have been possible. His involvement has been instrumental in ensuring the proper execution of our thesis work. We sincerely appreciate his contributions.

We are also grateful to **Mrs. Nafisa Sadaf**, Lecturer, Department of Computer Science & Engineering, IUT for her valuable inspection and suggestions on our proposal of data augmentation on low-level resources. We are truly thankful for her valuable opinions, dedication of time, and input provided, starting from the introduction of the research topics, selection of subjects, proposal of algorithms, modifications and implementation.

We are eternally thankful to **Dr. Md. Kamrul Hasan**, Professor, Department of Computer Science & Engineering, IUT for providing us guidance and suggestions throughout our research work. His knowledge and vast amount of insights and experience was key to the completion of our research work. We appreciate the effort he has put forward to guide us and provide his experienced opinion on our work. We are grateful to him for his contributions.

We would not have been able to complete this research without the guidance and assistance of several individuals who contributed in various ways. We would like to acknowledge and appreciate their valued help.

Contents

Declaration of Authorship	i
Abstract	iii
Acknowledgements	iv
List of Figures	vii
List of Tables	viii
1 Introduction	1
1.1 Problem Statement	3
1.2 Motivation and Scope	4
1.3 Research Challenges	7
1.4 Research Contribution	8
1.5 Thesis Outline	10
2 Literature Review	11
2.1 Data Augmentation	11
2.1.1 Data Augmentation Techniques	13
2.1.1.1 Rule-Based Techniques	13
2.1.1.2 Example Interpolation Techniques	14
2.1.1.3 Model-Based Techniques	15
2.2 Fake News Detection Datasets	16
2.2.1 Fake News Detection in Bengali	16
2.2.2 Bengali Fake News Datasets	17
2.2.3 Modern Fake News Detection Techniques	18
2.2.3.1 Deep Learning(DL) based methods	18
2.2.3.2 Attention Techniques	18
2.2.3.3 Transformer based Model	19
2.2.3.4 Others	19
2.3 Data Augmentation in Bengali	20
2.3.1 Limitations of Bengali Data Augmentation	21
2.3.2 Augmentation using BERT	22
2.3.3 Augmentation using Transfer Learning	23

2.3.4	Augmentation using ELECTRA and BanglaBert	24
2.3.5	Augmentation using N-gram with GRU and RNN	25
2.4	Evaluation using Semantic Similarity	26
2.4.1	Semantic Similarity in Texts	26
2.4.2	Semantic Textual Similarity in Data Augmentation	28
2.4.3	Semantic Similarity of Bengali Texts	28
3	Proposed Methodology	30
3.1	Architecture of Proposed Method	30
3.2	Model Description	31
3.3	Augmentation Procedure	31
4	Experimental Design	35
4.1	Evaluation Metric	35
4.2	Dataset Description	36
4.3	Experimental Setup	39
5	Results and Discussions	41
5.1	Experimental Result	41
5.2	Limitations	45
6	Conclusion and Future Work	46
	Bibliography	48

List of Figures

2.4	Augmentation using Transfer Learning [Adapted from [2]]	23
2.5	Augmentation using GRU and RNN [Adapted from [3]]	25
3.1	Architecture Overview	34
4.1	BanFakeNews Dataset Distribution	36
4.2	Imbalanced Augmented Dataset	37
4.3	Balanced Augmented Dataset	38
5.1	Imbalanced Dataset Loss with Dropout Layer but no class weights	41
5.2	Imbalanced Dataset Loss with Dropout Layer and Class Weights	42
5.3	Balanced Dataset Loss with Dropout Layer	42

List of Tables

3.1	Models Comparison	33
4.1	Dataset Evaluation Results	36
4.2	Dataset Distribution	38
4.3	Parameters of Our Model	39
4.4	Parameters of AugFakeBERT	40
4.5	Experimental Setup	40
5.1	Experimental Results	43
5.2	Results Comparison	43

*Dedicated to my parents, siblings, and friends for their continuous
encouragement of my academic endeavors and research . . .*

Chapter 1

Introduction

Natural Language Processing (NLP) is a field that lies at the intersection of artificial intelligence and computational linguistics [4]. The main emphasis is on creating algorithms and models that enhance the communication between computers and human language. The field of Natural Language Processing (NLP) aims to empower computers to comprehend, analyze, and produce human language in a way that is both meaningful and valuable, as shown in Figure 1.1 ².

²<https://www.javatpoint.com/nlp>

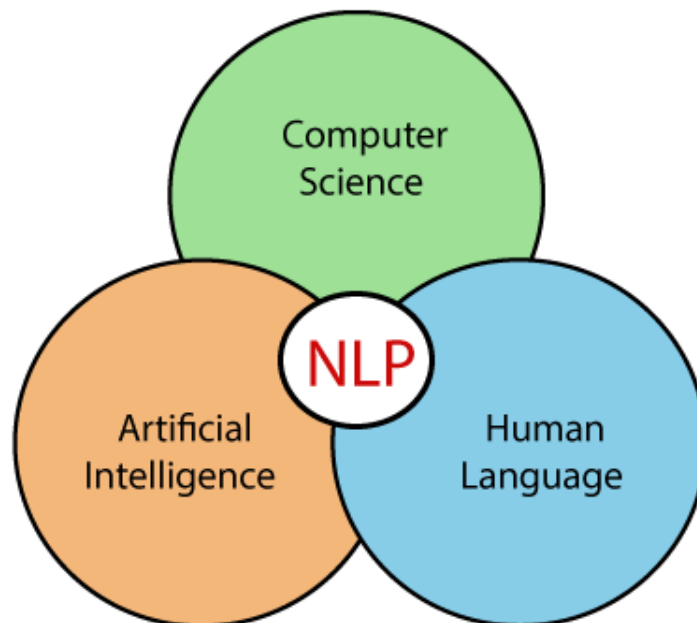


FIGURE 1.1: Natural Language Processing [Adapted from ¹]

Within the realm of NLP, a diverse range of tasks and applications exist [5]. These encompass the extraction of meaning and intent from text, including sentiment analysis, such as entity recognition, and text classification. NLP also encompasses machine translation, which translates speech/text from one language to another [6]. Furthermore, question answering systems are developed to comprehend questions in natural language and provide relevant and accurate responses.

NLP techniques typically involve a series of processes. These include tokenization, which breaks down large texts into smaller units such as words or subwords [7]. Parts-of-speech tagging is another process that assigns grammatical tags to words, enabling deeper analysis. Syntactic parsing is used to analyze sentence structure, while semantic analysis aims to extract meaning from text. Discourse analysis is employed to understand the context and coherence of text, enabling a more comprehensive understanding of language. Text generation is another crucial application of NLP, wherein the goal is to produce human-like text. This is particularly relevant in the development of chatbots, automated content creation, and language generation for storytelling purposes. Additionally, NLP involves information extraction, which entails identifying and extracting information that are structured from unstructured texts. This process can involve the extraction of entities, relationships, or events from textual data.

Fake news detection refers to the activity of locating and categorizing misinformation or intentionally false information disseminated through various media channels, including articles, newspapers, social media, and online platforms [8]. Fake news detection aims to distinguish between reliable, accurate information and fabricated or misleading content [9]. It's important to note that fake news detection using NLP is a complex and evolving field[10]. Combining multiple techniques, using robust labeled datasets, and continuously adapting models to evolving strategies employed by fake news producers can help improve the accuracy and effectiveness of fake news detection systems.

To address the challenges of detecting fake news using machine learning approaches, the availability of sufficient data becomes crucial[11]. However, in low-resource languages, obtaining a substantial amount of data is often challenging. Consequently, Data Augmentation emerges as a suitable technique to enhance the performance of models, albeit with careful consideration. Data augmentation refers to the artificial increase in the size of a dataset by generating additional data samples based on existing ones. This is achieved by applying diverse transformations or perturbations to the original data[12]. Examples of such transformations include rotating, shifting, or scaling the data, as well as introducing noise or other random perturbations. The primary motive of data augmentation is to introduce additional variations and diversity into the dataset. This

augmentation process helps to improve the performance and results of machine learning models by enhancing their robustness and capability to generalize unseen data.

By employing data augmentation techniques, models trained on limited data can benefit from a larger and more diverse dataset, thus potentially enhancing their performance [13]. However, it is crucial to exercise caution when applying data augmentation. Careful consideration should be given to the nature of the data, the specific characteristics of the low-resource language, and the potential impact of the augmentation techniques on the data's integrity and quality. Proper selection and application of data augmentation techniques can help mitigate the limitations posed by limited data availability in low-resource languages [14]. It is a constructive strategy to boost the performance of machine learning models in detecting fake news, making them more robust and capable of generalizing to new, unseen instances.

Semantic similarity in data augmentation refers to preserving the meaning or semantic content of the original text while generating new augmented samples [15]. When data augmentation techniques are applied on textual data, it is important to ensure that the augmented text remains semantically similar to the original text. In other words, the generated samples should convey similar meanings or intents as the original text. For example, if a data augmentation technique involves replacing certain words or phrases with synonyms, the goal is to choose synonyms that maintain the same/similar meaning as the original words [16]. This ensures that the augmented text retains the intended semantic information while introducing variation.

Preserving semantic similarity in data augmentation is crucial for maintaining the integrity and quality of the data [17]. If the augmented samples introduce substantial changes in meaning or distort the original semantics, it may lead to the generation of misleading or inaccurate data. This can negatively impact the performance of machine learning models trained on such augmented data. Overall, semantic similarity in data augmentation ensures that the augmented samples align with the intended meaning and semantics of the original text, thereby maintaining the coherence and integrity of the data.

1.1 Problem Statement

There are various problems with imbalance in data sets in low-resource languages. One of the interesting areas are on the domain of Fake News Detection models. In a low-resource language like Bengali only one published and reviewed data set exists geared towards this problem. [1] BanFakeNews is a data set for Fake News Classification problems. But

it suffers from a very big flaw which is a very high degree of imbalance in this data set for the classes.

One solution is to employ data augmentation techniques to generate augmented fake news texts based on existing fake news [18]. A key emphasis is placed on preserving semantic similarity between the augmented and original texts, ensuring coherence and maintaining the integrity of the augmentation process. By successfully augmenting the data, a balanced dataset comprising both fake and real news samples can be constructed. This balanced dataset serves as a benchmark for Bengali fake news detection. Subsequently, machine and deep learning models can be employed to evaluate the accuracy of the models using the test dataset. The objective is to validate that augmenting and balancing the dataset leads to improved accuracy compared to using an imbalanced and skewed dataset [19]. By leveraging the augmented and balanced dataset, the models can potentially achieve enhanced performance in detecting fake news in the Bengali language. The overall process involves two primary aspects: generating augmented fake news texts through data augmentation while preserving semantic similarity, and constructing a balanced dataset to serve as a benchmark for evaluating the performance of machine and deep learning models.

1.2 Motivation and Scope

The motivation to work on data augmentation and semantic similarity in Bengali lies in addressing the challenges posed by limited data, imbalanced datasets, and the need for improved generalization and robustness in detecting fake news. By employing these techniques, researchers and practitioners can become a contributor to the development of more accurate and reliable models for Bengali fake news detection [20], thereby combating the spread of misinformation and promoting trustworthy information dissemination in the language.

Our motivations are listed below:

- **Limited Data Availability:** Bengali, being a low-resource language, often faces a scarcity of labeled datasets for numerous natural language processing (NLP) tasks, including fake news detection. Data augmentation provides a means to artificially expand the available data, enabling the development and training of more robust models even with limited original data.
- **Unbalanced Datasets:** In the case of fake news detection in Bengali, there may be an imbalance between the number of real and fake news samples, making it

challenging to train accurate models [21]. Data augmentation, combined with techniques to preserve semantic similarity, can help balance the dataset by generating augmented samples of the underrepresented class (e.g., fake news). This leads to a more balanced and representative dataset, enhancing the model's performance.

- **Improved Generalization:** Data augmentation with semantic similarity preservation fosters better generalization of models to unseen data [22]. By generating augmented samples that retain the original semantic meaning, the models become more capable of handling variations and diverse instances in real-world scenarios, thereby improving their performance in detecting fake news.
- **Robustness against Variations:** By introducing diverse variations through data augmentation while maintaining semantic similarity, the models become more robust in handling different linguistic patterns, sentence structures, and expressions present in Bengali fake news. This robustness aids in capturing the subtle nuances and characteristics of fake news texts, enhancing the accuracy of detection.
- **Benchmark Dataset Creation:** Data augmentation, combined with semantic similarity preservation, can result in creating a dataset that is a benchmark for Bengali fake news detection. This balanced and augmented dataset can act as a valuable resource for evaluating the performance of different models, enabling fair comparisons and advancements in the field.

The scopes of working on data augmentation and semantic similarity in Bengali encompass improved fake news detection, the development of benchmark datasets, language-specific solutions, transferability to other low-resource languages, and addressing ethical implications related to misinformation [23]. These scopes contribute to the advancement of NLP techniques and support the creation of authentic and effective systems for detecting fake news in Bengali and beyond.

To explain them in detail, the scopes are:

- **Enhanced Fake News Detection:** The application of data augmentation and semantic similarity techniques in Bengali can significantly improve the accuracy and effectiveness of fake news detection models [11]. By generating augmented samples that maintain semantic coherence, the models can better capture the unique characteristics and linguistic nuances of Bengali fake news, leading to more reliable and precise detection.

- **Development of Benchmark Datasets:** Working on data augmentation and semantic similarity allows for the creation of benchmark datasets specifically tailored for Bengali fake news detection [24]. These datasets serve as standardized resources for evaluating the performance of different models and algorithms, fostering advancements in the field and enabling fair comparisons between different approaches.
- **Language-Specific Solutions:** Bengali is a distinct language with its own linguistic properties and challenges. By focusing on data augmentation and semantic similarity techniques specific to Bengali, researchers can develop language-specific solutions that cater to the unique characteristics and nuances of the language. This enables more accurate and contextually relevant fake news detection in Bengali.
- **Transferability to Other Low-Resource Languages:** The techniques and methodologies developed for data augmentation and semantic similarity in Bengali can be transferable to other low-resource languages facing similar challenges [25]. The insights gained from working on Bengali can serve as a foundation for addressing data scarcity and imbalance in other languages, contributing to the broader field of NLP in low-resource settings.
- **Ethical Implications:** Addressing fake news in any language, including Bengali, has significant societal and ethical implications. By improving the model accuracy of fake news detection, the spread of misinformation can be curbed, and the impact of fake news on public opinion and decision-making can be mitigated. The scopes of data augmentation and semantic similarity in Bengali extend to promoting information integrity and fostering a more informed society.

1.3 Research Challenges

Our primary research challenge was the limited availability of labeled data. Low-resource languages like Bengali generally suffers from a scarcity of labeled datasets, which restricts the scope of research and can impact the performance and generalization of models trained using data augmentation techniques [26]. Overcoming this challenge requires exploring strategies to effectively leverage small labeled datasets and developing techniques that can make the most out of the available data.

Semantic understanding of Bengali poses another significant research challenge. Bengali has its own linguistic nuances, idiomatic expressions, and cultural context, which make accurate semantic understanding a complex task [27]. Developing techniques that can capture the subtle semantic nuances and contextual variations in Bengali fake news texts requires in-depth understanding of the language and its unique characteristics. It involves delving into the intricacies of Bengali language usage and cultural references to ensure accurate interpretation and detection of fake news.

Preserving semantic coherence between the original and augmented texts is a critical challenge. Generating augmented texts that maintain semantic similarity without introducing distortions or misinterpretations is a complex task [28]. It necessitates careful selection of augmentation techniques, linguistic knowledge, and contextual understanding of the specific domain, such as fake news, in the Bengali language. Striking a balance between introducing variations and preserving semantic integrity requires innovative approaches and techniques.

Evaluating the effectiveness of different data augmentation techniques for Bengali fake news detection poses another challenge. Robust evaluation methodologies are required to determine the impact of various augmentation strategies on model performance and generalization [29]. Metrics such as accuracy, precision, recall, and F1 score need to be carefully considered. Establishing reliable evaluation protocols to assess the efficacy of different augmentation techniques is crucial to understand their true impact and identify the most effective approaches.

Preserving diversity in augmented data is a challenge that researchers must address [30]. While maintaining semantic similarity, it is important to ensure that the augmented dataset represents the diverse perspectives and variations present in Bengali fake news. Avoiding over-representation or bias towards specific linguistic patterns or topics is crucial. Curating a diverse and representative augmented dataset requires careful consideration and curation efforts.

Additionally, conducting research on fake news detection, including data augmentation and semantic similarity, carries ethical implications [31]. Adhering to ethical guidelines, safeguarding against unintended consequences, and promoting responsible use of technology is a critical challenge that researchers must address. Ethical considerations should guide the design, implementation, and evaluation of fake news detection models to ensure their responsible deployment and mitigate potential risks.

Addressing these research challenges necessitates interdisciplinary collaboration, linguistic expertise, data curation efforts, evaluation frameworks, and a comprehensive understanding of the unique characteristics of the Bengali language and the fake news domain. Overcoming these challenges will contribute to the advancement of fake news detection techniques and foster the development of more accurate and reliable systems for addressing the issue of fake news in the Bengali language.

1.4 Research Contribution

In “BanFakeNews: A Dataset for Detecting Fake News in Bangla” [1], the authors presented a dataset named “BanFakeNews” of 50,000 annotated news articles. Our contribution involves the improvement of a dataset that can be utilized for constructing automated fake news detection systems in low-resource languages like Bangla [32]. Given the issue of data imbalance within the dataset, we have made contributions by applying augmentation techniques that ensures semantic similarity. This augmentation process aims to enhance the dataset by balancing the distribution of fake news instances, thus improving the overall quality and effectiveness of the dataset for training fake news detection models.

The significant contributions of this thesis are listed below in brief:

- **Text Augmentation Technique and Dataset Improvement:** We want to create a well-balanced dataset specifically designed for training models in the detection of fake news in Bengali. This dataset should address the scarcity of existing datasets and the issue of highly unbalanced data. By collecting and curating a comprehensive dataset, researchers can provide a valuable resource for further studies in this domain.

We also want to develop a novel text augmentation technique tailored for the Bengali language. Since previous attempts at augmenting fake news texts in Bengali have yielded unsatisfactory results, there is a need for innovative approaches that can effectively preserve semantic relevance and structural accuracy. Developing

and implementing such techniques can significantly enhance the quality and diversity of training data.

- **Bengali Text Augmentation Framework:** Building upon the proposed framework, further contributions can involve refining and expanding the existing framework for fake news detection in Bengali [33]. This can include incorporating additional techniques, algorithms, or models to improve the overall effectiveness and reliability of the detection models. The framework can be iteratively improved and optimized based on empirical evaluations and feedback from real-world applications. The proposed framework aims to develop tailored evaluation metrics and benchmarks for fake news detection in Bengali, considering the language's unique characteristics and challenges. By establishing standardized evaluation protocols, researchers can facilitate fair comparisons between models [34]. Additionally, the framework holds potential for real-world application and deployment, allowing collaboration with media organizations, fact-checking platforms, and social media platforms to combat the spread of fake news and promote reliable information in the Bengali language community [35].

1.5 Thesis Outline

In Chapter 1, we have Introduced our topic, our ideas, motivation and scope, challenges and contribution in an elaborated manner. In Chapter 2, we have elaborated about the relevant background study and review of our research. In Chapter 3 , we discussed our proposed methodology with the dataset in detail. We focused on the model and algorithm we have applied to create the proposed framework. In Chapter 4 we explained all the experiments we have performed on the actual dataset and the new augmented dataset. Chapter 5 gives a comparative study of our work and results with the existing methods and models. Chapter 6 concludes the current investigation and outlines potential future initiatives. The final segment of this study contains all the references and credits used.

Chapter 2

Literature Review

The literature review consists of a detailed outline of the related research work in this field. Initially, it describes Data Augmentation and the techniques to text-based Data Augmentation. Then, the literature review focuses on fake news detection and its available datasets in Bengali Language. Then, it focuses on Data Augmentation in Bengali language and its limitations. Finally, the literature review concludes with explaining Semantic Similarity in Texts, how to relate semantic similarity with data augmentation and the related work of semantic similarity in the bengali language.

2.1 Data Augmentation

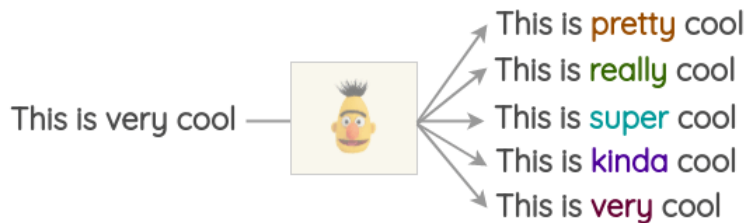


FIGURE 2.1: Example of Text Augmentation [Adapted from ¹]

Data augmentation in text is the process of generating new, synthetic text samples by applying various transformations to the existing text data [36]. The goal is to expand the size and diversity of the training dataset, which can improve the performance and robustness of machine learning models as shown in 2.1 ². There are several techniques used for data augmentation in text. One common approach is to apply variations to the text itself, like replacing words with their synonyms, or inserting or deleting words, or

²<https://www.analyticsvidhya.com/blog/2022/02/text-data-augmentation-in-natural-language-processing-v>

shuffling the word order. These techniques can introduce different textual representations while preserving the original meaning and context.

Another technique involves incorporating linguistic rules or knowledge into the augmentation process. For example, grammatical transformations like tense changes, negation, or passive voice conversion can be applied to the text[37]. These transformations adhere to the linguistic structure and rules of the language, resulting in semantically meaningful augmented samples as shown in Figure 2.1³. Furthermore, data augmentation can also involve external resources, such as translation or paraphrasing models. Translating the text into a different language and then translating it back to the original language can introduce new word choices and expressions. Paraphrasing techniques can be used to rephrase sentences while maintaining the same underlying meaning.

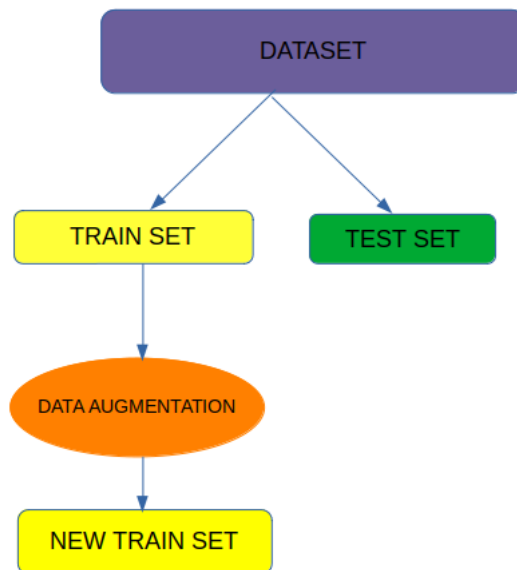


FIGURE 2.2: Text Augmentation [Adapted from⁴]

Data augmentation in text has several benefits. It helps address the challenge of limited training data, particularly in low-resource languages or specialized domains. Augmented data can enhance model generalization by exposing the model to a wider range of variations and patterns[38]. It can also alleviate issues like overfitting and imbalanced datasets by balancing class distributions and increasing the diversity of samples. However, it is important to consider certain factors when applying data augmentation in text. The augmented samples should retain their semantic integrity and coherence to ensure they

³<https://neptune.ai/blog/data-augmentation-nlp>

are still meaningful and representative of the original data. Care must be taken to avoid introducing biases or distortions in the augmented dataset[12]. Evaluating the impact of different augmentation techniques on model performance and selecting appropriate augmentation strategies are key considerations.

2.1.1 Data Augmentation Techniques

2.1.1.1 Rule-Based Techniques

One notable example of a rule-based technique in text data augmentation is EDA (Easy Data Augmentation), as described in the paper by Wei and Zou titled “EDA: Easy Data Augmentation for Boosting Performance on Text Classification Tasks” (2019) [39]. EDA is a technique specifically designed to enhance the performance of text classification tasks. It involves four operations: replacement of synonyms, inserting randomly, swapping randomly, and deleting randomly. These operations are applied to augment the training data and increase its diversity.

Synonym replacement involves replacing words in the text with their synonyms while preserving the overall meaning. This helps introduce variations in word choice and expression [40]. For example, the word “good” might be replaced with “excellent” or “superb.” Random insertion involves adding new words into the text at random positions. These additional words can introduce additional context or expand the vocabulary representation. For instance, a sentence like “The cat is sitting on the mat” might be augmented by adding the word “comfortably” to become “The cat is sitting comfortably on the mat” as seen in Figure 2.1.1.1 ⁵.

Challenge of Semantically Invariant Transformation in NLP

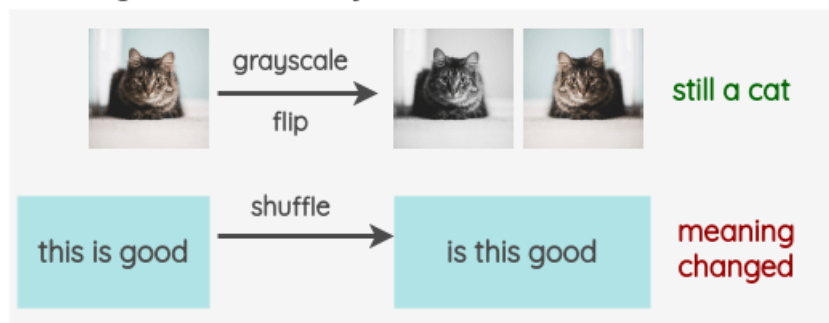


FIGURE 2.3: Limitations of Text Augmentation [Adapted from ⁶]

⁵<https://paperswithcode.com/task/text-augmentation>

Random swap entails exchanging the positions of two words within the text. This operation helps capture different word orderings and syntactic variations. For instance, in the sentence “I love ice cream,” the words “love” and “ice cream” could be swapped to create “I ice cream love”. Random deletion involves removing words from the text at random. This operation can simulate missing or incomplete information and enhance the ability of the model to handle noisy or truncated text samples. For example, the sentence “He is going to the store” might be augmented by deleting the word “going” to become “He is to the store.”

By applying these four operations, EDA facilitates the creation of augmented text samples that exhibit variations in word usage, word order, and sentence structure [12]. This increased diversity in the training data can improved model performance and generalization, particularly in text classification tasks. The EDA technique is simple to implement and has demonstrated its effectiveness in various text classification scenarios. It offers a practical and efficient way to augment text data and enhance the robustness of models in handling different variations of text inputs.

2.1.1.2 Example Interpolation Techniques

The technique referred to as MIXUP, as introduced in the paper by Zhang et al. titled “mixup: Beyond Empirical Risk Minimization” (2017), involves the interpolation of inputs and labels from multiple real examples. This approach, also known as Mixed Sample Data Augmentation (MSDA), was initially used primarily for tasks involving continuous inputs [41]. Traditionally, MIXUP was applied by taking two or more input examples, combining them through linear interpolation, and generating new augmented examples. The same interpolation was performed for the corresponding labels, resulting in augmented label distributions. By blending the inputs and labels, MIXUP aimed to motivate the model in learning more general and robust decision boundaries [42].

However, there was a limitation with MIXUP when it came to tasks that involved non-continuous inputs, such as text data [41]. To overcome this limitation, recent advancements have extended MIXUP to work with embeddings or higher hidden layers. In particular, the paper by Chen et al. titled “MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification” (2020) proposed a method called MixText [43], which enables the application of MIXUP in the context of text data [44].

In MixText, the interpolation is performed in the hidden space, such as embeddings or higher layers of a neural network, rather than directly on the input data [45]. This allows

for the mixing of text representations, leveraging linguistic information, and generating augmented samples. By mixing the hidden representations, MixText enhances the ability of the model to handle variations in text inputs and improves its performance in semi-supervised text classification tasks. The extension of MIXUP to text data through techniques like MixText has expanded the applicability of this data augmentation approach beyond tasks with continuous inputs. It enables the generation of augmented text samples by interpolating hidden representations, providing a valuable tool for enhancing the performance and generalization of models in text-related tasks.

2.1.1.3 Model-Based Techniques

The BACKTRANSLATION method, introduced in the paper by Sennrich et al. titled “Improving Neural Machine Translation Models with Monolingual Data” (2015), is a data augmentation technique that involves translating a sequence of text into another language and then translating it back into the original language [46] [47]. This process helps generate new augmented examples by introducing variations in word choice and sentence structure. By leveraging the translations, the BACKTRANSLATION method aims to increase the quantity and the diversity of training data, ultimately improving model performance.

Another data augmentation technique is described in the paper by Kobayashi et al. titled “Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations” (2018) [48] [47]. The approach, known as contextual augmentation, entails substituting words in the text with randomly selected alternatives determined by the distribution of a recurrent language model in the given context. By incorporating contextual information, this technique generates augmented examples that exhibit variations while preserving the overall meaning and coherence of the text.

In the paper by Yang et al. titled “Generative Data Augmentation for Commonsense Reasoning” (2020) [49], a method called GDAUG is proposed. GDAUG utilizes pre-trained transformer language models to generate synthetic examples for data augmentation. It chooses the most informative and diverse set of augmented examples based on certain criteria, such as their relevance to the task or their ability to cover a broad range of input variations. By incorporating generative models, GDAUG aims to create high-quality synthetic examples that can enhance model training and generalization.

These data augmentation methods, including BACKTRANSLATION, contextual augmentation, and GDAUG, provide techniques for generating augmented examples in order

to enrich the training data and improve machine learning models's performances. By leveraging language translation, contextual word replacement, and generative models, these methods offer ways to introduce variations, increase diversity, and enhance the robustness of models in various natural language processing tasks.

2.2 Fake News Detection Datasets

Fake news are articles that have the potential to misdirect readers by presenting false or incorrect information. The primary purpose of fake news is often to tarnish the reputation of any entity, or individual. The proliferation of social media platforms like Facebook, coupled with the ease of accessing online advertising revenue, has contributed to the widespread dissemination of fake news [50]. Additionally, the increased political polarization within society has further fueled the spread of misinformation. Furthermore, there have been instances where government actors have been set up in generating and disseminating fake news in a hostile manner, usually during elections [9] [1]. Their objective is to manipulate individuals for personal or organizational gains. This manipulation through fake news has become a growing concern in recent years.

The 2012 Ramu incident in Bangladesh serves as a notable example, highlighting the destructive consequences that can result from the spread of fake news⁷. In this incident, approximately 25 thousand individuals took part in the destruction of Buddhist temples based on a Facebook post originating from a fake account. The enraged mob managed to destroy around 12 Buddhist temples and monasteries, as well as 50 houses. This incident underscores the potential for fake news containing blasphemous content to incite similar acts of violence, particularly in cases where people hold strong religious sentiments. To address the issue of fake news, there are websites that are dedicated to provide such news like www.politifact.com, www.factcheck.org, and www.jaachai.com. These platforms manually curate and update potential fake news stories that are published online, providing logical and factual explanations to debunk false information. However, these websites have limitations, as they are not equipped to respond swiftly to emerging fake news events. The timely detection and debunking of fake news remain ongoing challenges that require more efficient and agile solutions[51].

2.2.1 Fake News Detection in Bengali

In recent times, computational approaches have emerged as valuable tools in combating the spread of fake news. Long et al.(2017)[52] explored the use of multi-perspective

⁷<https://www.thedailystar.net/backpage/news/eight-years-ramu-attack-buddhists-still-wait-justice-196>

speaker profiles to detect fake news, while Yang et al. (2017)[53] focused on utilizing linguistic features for detecting satirical news. Karadzhov et al. (2017)[54] proposed a fact-checking model that is automated fully and relies on external sources to verify the claims made in news stories. Dong et al. (2019)[55] employed deep two-path semi-supervised learning to identify news in social media that are fake. However, these studies have predominantly focused on news published in English.

It is worth highlighting that Bengali, with approximately 341 million speakers, is the fifth most widely spoken language in the world [56]. Surprisingly, despite the significant number of Bengali speakers, there is a huge current lack of resources and computational approaches dedicated to addressing the risk of fake news in the Bengali language. This absence of resources and tools specifically tailored to tackle fake news in Bengali has the potential to negatively impact this substantial population. Therefore, there is a pressing need for resource development and computational approaches to mitigate the risks associated with fake news in the Bengali language and safeguard the interests of this large group of individuals.

2.2.2 Bengali Fake News Datasets

The first Bengali fake news dataset was published on April 2020 which consisted of 50,000 data where 48,678 are real news and 1,299 are fake news [1]. In order to gather a collection of genuine news articles, they chose 22 widely recognized and widely read news portals in Bangladesh. To gather false news articles, they included the following categories [1]:

- **Misleading/False Context:** Any news that includes unreliable information or presents facts that can mislead readers.
- **Clickbait:** Any news that includes unreliable information or presents facts that can mislead readers.
- **Satire/Parody:** News stories that are created for entertainment and parody purposes

The assessment of linear classifiers and neural network-based models indicates that traditional linguistic features combined with linear classifiers may outperform neural network-based models [57]. Notably, character-level features are found to hold greater significance compared to word-level features. Furthermore, the analysis reveals that the occurrence of punctuations is more prevalent in fake news compared to authentic news.

Additionally, it is observed that fake news predominantly originates from less popular websites.

2.2.3 Modern Fake News Detection Techniques

2.2.3.1 Deep Learning(DL) based methods

Deep Learning techniques, such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN), have been shown to be effective in identifying false news.[58] developed an optimized CNN (OPCNN-FAKE) model for detecting fake news, and their findings showed that this model, with optimized parameters, performed better than other models including RNN and traditional classifiers. Ensemble approaches, which combine multiple models, are commonly used in this field and often involve the use of both CNN and RNN. CNN is used for feature extraction, while long short-term memory (LSTM) is used for classification [59] [60]. While these approaches have proven successful, there is room for improvement, and using contextual information from textual content features can lead to better results.

2.2.3.2 Attention Techniques

Recently, attention techniques have been used to effectively extract information from news articles in relation to a mini query.[61] developed a three-level hierarchical attention network (3HAN) that uses attention at the word, sentence, and headline level to assign weights to different parts of an article. 3HAN has been shown to produce more accurate results than other deep learning models, even though it only uses textual data. However, many previous neural network approaches have used non-contextual embedding, which does not consider the contextual meaning of the text and has therefore had limited success. In addition, many models struggle to achieve satisfactory performance in detecting fake news. To address these issues, [62] proposed the self-attention-based Automatic Fake News Classification (ACT) method, which relies on the mutual interaction between a claim and supporting replies. ACT uses self-attention in combination with the LSTM model to extract key components of an article using multiple feature vectors, considering the internal correlation between words in the article. Another approach, called Graph-aware Co-Attention Networks (GCAN)[63] , has been developed to detect fake news on short-text Twitter posts based on the historical record of its retweeters. However, this model is not suitable for long texts.

2.2.3.3 Transformer based Model

Another development in the field of natural language processing is the use of language models, which produce contextual embeddings, for fake news detection tasks. Bidirectional Encoder Representation (BERT) is a commonly used language model in this context.[64] improved the performance of BERT in fake news detection by including news data in the pre-training phase. The resulting model, called ExBAKE (BERT with additional unlabeled news text data), outperformed the state-of-the-art stackLSTM model by a 0.137 F1 score.[65] proposed a BERT-based Domain-Adaption Neural Network (BDANN) for multimodal fake news detection, which uses BERT to extract text features and a pre-trained VGG-19 CNN model to extract visual features. These features are then combined and used by the detector to distinguish between fake and real news. However, the presence of corrupted image data in the Weibo dataset affected the results of BDANN.[66] developed a fake news detection mechanism called fakeBERT, which combines BERT with a deep convolutional approach using a 1D-CNN with varying kernel sizes and filters. This combination can handle both large-scale structured and unstructured text, making it effective in dealing with ambiguity.

2.2.3.4 Others

Wu et al.[67] proposed a multi-modal framework that combines RNN and Graph Convolutional Network (GCN) for encoding fake news propagation on social media. One issue in the field of natural language processing is the limited data available on fake news, and some studies have attempted to generate synthetic data to address this problem.[68] introduced the Sequence Generative Adversarial Network (SeqGAN), a GAN structure that uses a reinforcement learning-oriented approach and Monte Carlo search to solve the gradient descent problem in GANs for discrete outputs [2]. The authors fed real news content into the GAN and then trained a classifier based on Google's BERT model to distinguish actual samples from generated samples [2]. However, SeqGAN is not suitable for generating entire news content and may produce non-contextual news.[68] also used DistilBERT to encode the representation of textual data such as fake news, tweets, and user self-descriptions, and improved the performance by incorporating an attention mechanism. Limited data on fake news remains a significant issue in the field of natural language processing, and generating synthetic data may not be a complete solution.

2.3 Data Augmentation in Bengali

Data augmentation in the Bengali language involves applying various techniques to generate additional augmented examples and increase the diversity of the training data specifically for Bengali text [69]. These techniques aim to improve the performance and generalization of machine learning models in Bengali language-related tasks, such as text classification, sentiment analysis, machine translation, and more. Some commonly used data augmentation techniques in Bengali language include:

- **Synonym Replacement:** This technique involves replacing words in Bengali text with their synonyms while preserving the overall meaning [40]. By incorporating Bengali language resources such as lexical databases or word embeddings, alternative words can be substituted, introducing variations in word choice.
- **Transliteration:** Transliteration involves converting Bengali text into its phonetic or romanized representation [70]. This technique can be useful in tasks where handling romanized text is beneficial, such as named entity recognition or speech processing. Transliteration can help augment the data by generating additional examples in a different representation.
- **Contextual Word Replacement:** Similar to the contextual augmentation technique mentioned earlier, Bengali-specific contextual word replacement involves replacing words in Bengali text based on the distribution of a language model within the current context [71]. This technique helps create augmented examples while preserving the semantic meaning and coherence of the text.
- **Text Transformation:** Text transformation techniques involve applying modifications to the structure or form of Bengali text [72]. This can include tasks such as text summarization, paraphrasing, or sentence splitting, where the original text is transformed into a modified version. These transformations can introduce variations and increase the diversity of the training data.
- **Backtranslation:** Backtranslation, as mentioned earlier, involves translating Bengali text into another language and then translating it back into Bengali [73]. This technique leverages machine translation systems to generate augmented examples with variations in word choice and sentence structure.

It is important to adapt these data augmentation techniques specifically for the Bengali language, taking into consideration the linguistic and cultural characteristics of the language. Additionally, the availability of Bengali language resources, such as lexical

databases, word embeddings, or pretrained language models, plays a crucial role in implementing effective data augmentation techniques for Bengali text.

2.3.1 Limitations of Bengali Data Augmentation

Data augmentation in the Bengali language presents several challenges due to the specific characteristics of the language and the availability of resources. One major difficulty is the limited availability of language resources. Unlike widely spoken languages such as English, Bengali has fewer linguistic resources, including lexicons, word embeddings, and pretrained language models [74]. This scarcity of resources can hinder the effectiveness and diversity of data augmentation techniques in Bengali.

Another challenge is the lack of large labeled datasets in Bengali. Data augmentation techniques often rely on a substantial amount of labeled data for training and generating augmented examples. However, in the case of Bengali, there may be a scarcity of large labeled datasets for specific tasks. The limited availability of data can restrict the effectiveness of data augmentation techniques and hinder the overall performance of models.

Bengali has its own linguistic complexities, including intricate grammar, rich morphology, and unique word order [74]. These language-specific challenges make it difficult to directly apply data augmentation techniques that are primarily designed for languages with different linguistic properties. Adjustments and adaptations are required to ensure that the generated augmented examples maintain linguistic integrity and coherence in Bengali.

Cultural and contextual sensitivity is another aspect to consider in data augmentation for Bengali. Bengali text often contains cultural references, idiomatic expressions, and domain-specific vocabulary [75]. Preserving the cultural and contextual sensitivity of the language is crucial during augmentation to ensure that the generated examples are relevant, contextually accurate, and respectful of Bengali cultural nuances.

Additionally, domain-specific challenges may arise in data augmentation for Bengali, depending on the task or domain at hand. Specialized domains like healthcare, legal, or finance may require domain-specific knowledge and expertise to effectively augment the data while maintaining the integrity and relevance of the content. Augmenting Bengali text in these domains requires careful consideration of the specific terminology, jargon, and linguistic conventions associated with each domain.

To address these difficulties, researchers and practitioners working on data augmentation in Bengali need to focus on developing language-specific resources, creating domain-specific datasets, and exploring innovative techniques that consider the unique linguistic and cultural characteristics of Bengali [76]. Collaboration among researchers, the creation of annotated datasets, and the development of language models specifically for Bengali can contribute to overcoming the challenges and advancing data augmentation techniques in the language.

2.3.2 Augmentation using BERT

Text Augmentation can be done by utilizing a Bidirectional Encoder Representation of Transformers (BERT) language model to generate an augmented dataset consisting of synthetic fake data. BERT is a powerful language model that has demonstrated impressive performance in various natural language processing tasks [77]. By leveraging BERT, the proposed approach aims to generate synthetic fake data that can be used to augment the existing dataset. This augmentation technique helps overcome the issue of having an imbalanced dataset, where the minority class (in this case, fake data) may be underrepresented.

BERT consists of two main stages: pre-training and fine-tuning. In the pre-training stage, BERT is exposed to a vast amount of unlabeled data and learns to perform a variety of language understanding tasks [77]. This enables the model to acquire a deep bidirectional representation of words and sentences, capturing the contextual relationships between them. After the pre-training stage, BERT is fine-tuned using labeled data specific to a particular task. This labeled data is typically related to a specific natural language processing task, such as question answering or sentiment analysis. The fine-tuning process involves training the model on the task-specific data, allowing it to adapt its parameters to the specific nuances and patterns of that task.

One notable advantage of BERT is its ability to be fine-tuned for various tasks with minimal changes to the model architecture. This means that the same pre-trained BERT model can be used as a starting point for different tasks, saving significant computational resources and time. During the pre-training stage, BERT is trained on two specific tasks. The first task is the Masked Language Model (MLM), where a certain percentage of the input tokens are randomly masked, and the model is trained to predict the original tokens based on the context provided by the surrounding words [78]. This MLM task allows BERT to learn the contextual relationships between words and improve its understanding of the overall text.

The second task BERT is trained on is Next Sentence Prediction (NSP). In this task, BERT learns to predict whether one sentence follows another in a given corpus [79]. By training on this task, BERT gains the ability to understand the relationships between sentences and capture the discourse and coherence within a text. Overall, BERT’s pre-training and fine-tuning stages enable it to learn powerful representations of words and sentences, capturing the contextual information necessary for a wide range of natural language processing tasks. Its flexibility in fine-tuning for different tasks makes it a highly versatile and effective model in the field of natural language processing.

2.3.3 Augmentation using Transfer Learning

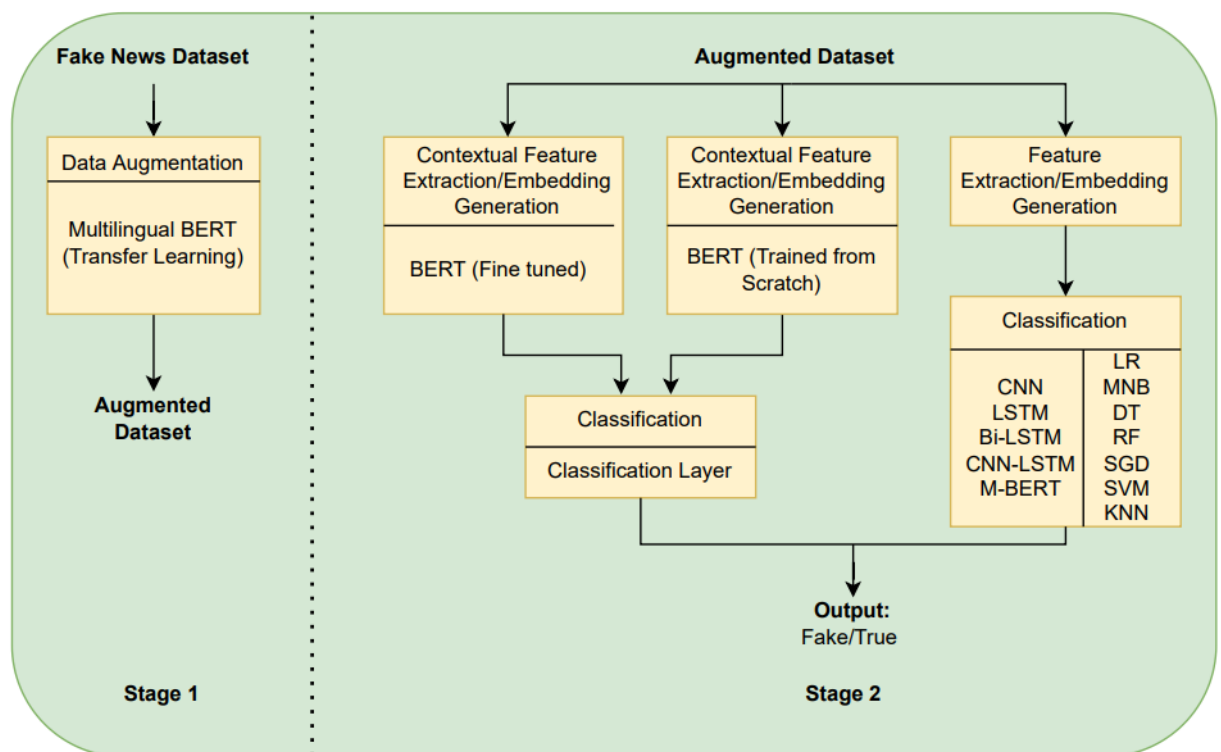


FIGURE 2.4: Augmentation using Transfer Learning [Adapted from [2]]

Performing Bengali text augmentation using transfer learning entails utilizing pre-trained language models that have been trained on extensive text data as shown in Figure 2.3.3, specifically in Bengali or in multilingual contexts encompassing Bengali [80]. The process can be delineated into several sequential steps. Initially, it is necessary to acquire a suitable pre-trained language model that encompasses the linguistic characteristics of Bengali. These models have undergone training on vast text corpora and have acquired contextual representations of words and phrases, rendering them effective for diverse language-related tasks. Subsequently, the pre-trained language model must be

fine-tuned using the specific Bengali text dataset at hand. Fine-tuning involves training the model on the target domain, enabling it to adapt to the particular nuances and patterns inherent in the data. Techniques such as masked language modeling or next sentence prediction can be employed to refine the model’s comprehension of the Bengali language.

Once the language model has been fine-tuned, it can be leveraged to generate augmented text samples. Various techniques can be employed to introduce variations into the original text. Synonym replacement, backtranslation, character-level perturbations, or contextual word replacement can be employed to create augmented versions of the original text. It is of paramount importance to ensure that the augmented text maintains semantic coherence and integrity [81]. The introduced variations should uphold the meaning and context of the original text while incorporating diversity. This ensures that the augmented data remains faithful to the characteristics of the Bengali language and preserves the intended meaning.

The subsequent step involves merging the original dataset with the augmented data. By combining both datasets, a larger and more diverse dataset is created, facilitating the training of models. This augmented dataset can enhance the performance and robustness of the models by providing additional variations and examples for learning. Finally, the machine learning or deep learning models can be trained or fine-tuned using the augmented dataset. The models can now benefit from learning patterns and representations from both the original and augmented data, enabling them to generalize better and effectively handle variations in Bengali text.

2.3.4 Augmentation using ELECTRA and BanglaBert

ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) [82] is a pre-training method introduced as an alternative to the popular masked language modeling (MLM) approach used in models like BERT [82]. While MLM corrupts input by replacing tokens with the special [MASK] token, ELECTRA employs a different technique called replaced token detection (RTD).

In the RTD approach, input tokens are corrupted by replacing them with plausible alternatives generated by a small network known as the generator [83]. The corrupted input is then used to train a discriminative model, the discriminator, to predict which tokens in the corrupted input have been replaced. This method differs from MLM in that it operates on all input tokens rather than a subset, resulting in a more sample-efficient pre-training approach.

ELECTRA has demonstrated several advantages over MLM. Firstly, it produces contextual representations that outperform those of BERT when using the same model size, data, and computational resources. This is because RTD leverages the information from all tokens in the input, leading to more effective learning [84]. Secondly, ELECTRA achieves comparable performance to other state-of-the-art models like RoBERTa and XLNet while requiring less computational resources, making it more efficient at scale.

In the context of BanglaBERT, which is a pre-trained model specifically designed for the Bengali language, ELECTRA is utilized as the underlying pre-training method [85]. By employing the RTD approach with a generator and discriminator model, BanglaBERT benefits from the improved performance that comes with leveraging information from all tokens in a sequence. Compared to MLM, which only uses 15% of the tokens for training, RTD allows for better performance. Moreover, ELECTRA has shown promising results similar to those of RoBERTa [86] and XLNet [87], but with reduced training time, making it a suitable choice for implementing BanglaBERT.

2.3.5 Augmentation using N-gram with GRU and RNN

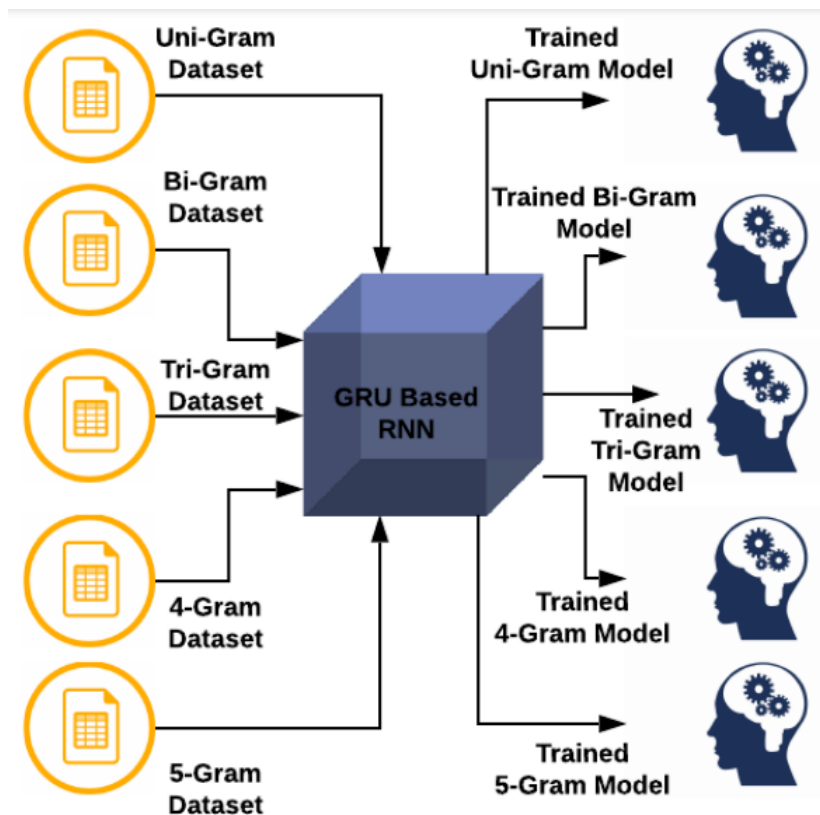


FIGURE 2.5: Augmentation using GRU and RNN [Adapted from [3]]

Text prediction systems are designed to enhance the efficiency of textual communication by suggesting the most likely word to follow in a given language. In this particular study, a method is proposed for predicting the next appropriate word in the Bengali language. The approach utilizes a Gated Recurrent Unit (GRU)-based Recurrent Neural Network (RNN) trained on an n-gram dataset as shown in Figure 2.3.5 [88] [89].

To conduct the study, a corpus dataset is collected from various sources in Bengali. The method is then compared against other techniques such as an RNN based on Long Short-Term Memory (LSTM) architecture using n-grams, as well as Naive Bayes with Latent Semantic Analysis [90]. The results of the study demonstrate the effectiveness of the proposed method. It outperforms the LSTM-based RNN on n-grams and the Naive Bayes approach with Latent Semantic Analysis [91]. The average accuracy achieved by the proposed method is reported as 99.70% for 5-grams, 99.24% for 4-grams, 95.84% for 3-grams, 78.15% for 2-grams, and 32.17% for 1-grams.

These results highlight the strong predictive capabilities of the GRU-based RNN approach for Bengali text prediction [3]. The high accuracy across different n-gram sizes indicates its proficiency in suggesting the most appropriate next word in Bengali language contexts.

2.4 Evaluation using Semantic Similarity

2.4.1 Semantic Similarity in Texts

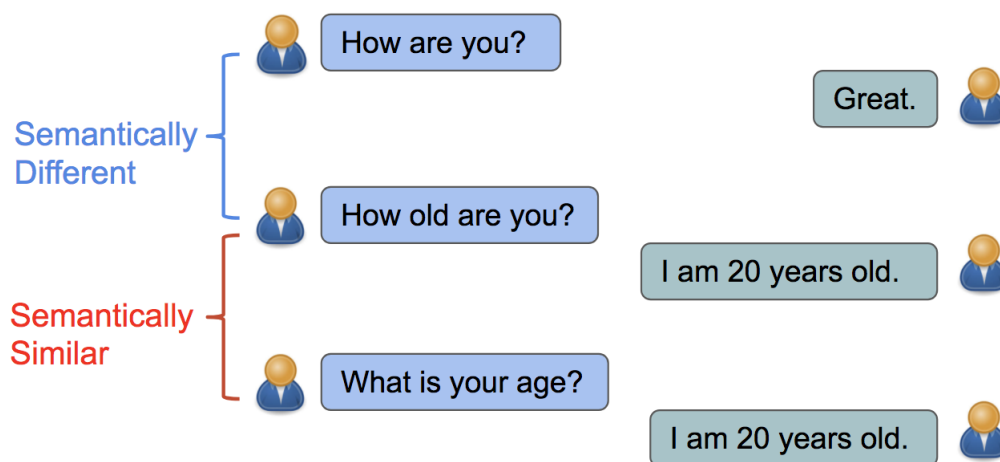


FIGURE 2.6: Semantic Similarity [Adapted from 8]

Semantic similarity in text refers to the degree of resemblance or similarity between two or more pieces of text in terms of their meaning or semantic content⁹. It focuses on capturing the similarity or relatedness of the underlying concepts, ideas, or information conveyed by the texts, rather than just considering the surface-level similarity based on words or phrases as shown in Figure 2.4.1. In natural language processing (NLP), measuring semantic similarity is important for various tasks such as information retrieval, text summarization, question answering, and language generation. It allows systems to understand the similarity between different texts, identify relevant documents or passages, and generate coherent and contextually appropriate responses.

There are several computational methods and algorithms used to assess semantic similarity in text. These methods aim to capture the semantic relationships between words, phrases, sentences, or even entire documents [92]. They consider various linguistic features, such as word meanings, syntactic structures, semantic relationships, and contextual information, to determine the similarity between texts. Common techniques for computing semantic similarity include:

- **Word embeddings:** Representing words as dense vectors in a high-dimensional space, where similar words are located closer to each other.
- **Semantic networks:** Utilizing knowledge graphs or ontologies to capture the semantic relationships between concepts or entities.
- **Supervised and unsupervised learning approaches:** Training models to learn the similarity based on labeled or unlabeled data.
- **Deep learning models:** Using neural networks, such as recurrent neural networks (RNNs) or transformer models, to capture the contextual and semantic information in text.

The output of semantic similarity models is typically a similarity score or a distance metric that quantifies the degree of similarity between two texts. This score can range from 0 (completely dissimilar) to 1 (identical or highly similar), or it can be a continuous value that represents the degree of similarity [93]. Overall, semantic similarity in text is a fundamental concept in NLP that enables machines to understand and compare the meaning of textual content. It plays a vital role in various language understanding and generation tasks, enhancing the performance and accuracy of NLP systems.

⁹<https://ai.googleblog.com/2018/05/advances-in-semantic-textual-similarity.html>

2.4.2 Semantic Textual Similarity in Data Augmentation

Semantic similarity is important in data augmentation because it helps ensure that the augmented data retains the same meaning and context as the original data [94]. When performing data augmentation, the goal is to create new synthetic examples that are similar to the original examples but introduce additional variations. By considering semantic similarity during data augmentation, we can generate new samples that preserve the underlying semantics of the text [95]. This is crucial because simply making random changes to the text without considering its meaning can lead to the generation of nonsensical or incorrect samples. Semantic similarity in data augmentation helps in the following ways:

- **Preserving Meaning:** By maintaining semantic similarity, the augmented data remains coherent and retains the same overall meaning as the original data. This ensures that the augmented samples are still relevant to the task or problem at hand.
- **Maintaining Integrity:** Augmented samples with high semantic similarity maintain the integrity of the data [96]. They do not introduce misleading or contradictory information that may adversely affect the performance of models trained on the augmented data.
- **Generalization:** Augmenting data with semantic similarity helps models generalize better. If the augmented data captures the variations and diversity present in the original data while maintaining semantic similarity, models trained on this augmented data are more likely to perform well on unseen or real-world data.
- **Dataset Balancing:** Semantic similarity can also be useful in balancing datasets during data augmentation. By ensuring that the augmented samples are semantically similar to the original samples, we can generate balanced datasets that represent different classes or categories in a more equitable manner.

Overall, incorporating semantic similarity in data augmentation techniques improves the quality and usefulness of the augmented data [97]. It helps maintain the original intent and meaning of the text, leading to more effective training of machine learning models and better performance on downstream tasks.

2.4.3 Semantic Similarity of Bengali Texts

In the field of information retrieval, determining the semantic similarity between texts plays a vital role in various applications. However, when it comes to Bengali language,

there is a lack of widely available methods for measuring semantic similarity . In a research paper by Shajalal et al. titled “Semantic Similarity Measurement of Bengali Texts Using Pre-trained Word Embeddings” [98], an approach is proposed to estimate the semantic similarity between different segments of Bengali text. The researchers leverage a pre-trained word embedding model, which is a type of language representation model that captures the semantic relationships between words.

The proposed approach involves applying a specific algorithm that utilizes the pre-trained word embeddings to measure the semantic similarity between Bengali text segments [99]. The algorithm takes into account the contextual meaning and relationships of words within the text. To evaluate the effectiveness of their approach, the researchers conducted experiments using a dataset consisting of Bengali texts. They compared their method’s performance against existing techniques for measuring semantic textual similarity in Bengali.

The results of the study demonstrated that the proposed approach achieved state-of-the-art performance in terms of the Pearson correlation coefficient, which is a commonly used metric for assessing the correlation between two variables. This indicates that the method successfully captured the semantic similarity between Bengali text segments and produced reliable similarity scores. By introducing an effective method for measuring semantic similarity in Bengali texts, the research contributes to the field of natural language processing in the Bengali language. It provides a valuable tool for various information retrieval applications that rely on semantic understanding and similarity calculations in Bengali text data.

Chapter 3

Proposed Methodology

3.1 Architecture of Proposed Method

Our model, the mT5-small model, is derived from the BanglaParaphrase[100] dataset, which is a high-quality Bangla paraphrase dataset. The mT5-small model is a specific variant of the mT5 model, which, in turn, is the multilingual variant of the T5 (Text-To-Text Transfer Transformer) model[101]. The mT5 model is trained on a diverse dataset consisting of 101 languages, including Bangla.

The mT5 model incorporates the advancements of the T5 model[102] and is specifically designed for multilingual natural language processing tasks. It leverages pre-training on a large-scale Common Crawl-based dataset to learn language representations and capture language patterns across various languages, including Bangla.

In our research, we utilize the mT5-small model, which is a smaller and more lightweight variant of the mT5 model. The mT5-small model maintains the capabilities of the larger model while offering improved efficiency and faster inference times.

By leveraging the mT5-small model and combining it with the filtering methods outlined in BanglaParaphrase, we are able to generate new fake news data by paraphrasing existing samples. This process ensures that the generated data maintains the original meaning and intent of the news articles while providing diverse variations.

Overall, the mT5-small model, derived from BanglaParaphrase, serves as a powerful tool for generating new fake news data by leveraging its pre-training on a diverse multilingual dataset and the paraphrasing techniques incorporated in BanglaParaphrase. This model enables us to effectively expand and enhance the dataset for our research purposes.

3.2 Model Description

Our model the mT5-small model taken from BanglaParaphrase: A High-Quality Bangla Paraphrase Dataset[100]. mT5-small is a variant of the mT5[101] model which in turn is the multilingual variant of the T5[102] model. mT5 is a multilingual variant of T5 that was pre-trained on a new Common Crawl-based dataset covering 101 languages. It is used together with the filtering methods in BanglaParaphrase[100] in order to generate the new fake news data by paraphrasing.

3.3 Augmentation Procedure

Nowe present our Augmentation Procedure for augmenting the fake news dataset and handling the class imbalance issue. The methodology consists of several steps, including obtaining original fake news headlines, employing a pre-trained Bangla T5 model for paraphrasing, preprocessing and filtering the headlines, generating additional paraphrased versions, and combining the original and generated headlines. The aim is to expand and diversify the dataset while maintaining the integrity of the original samples. Additionally, we discuss strategies to address the class imbalance, such as undersampling the majority class (real news) to achieve a balanced dataset. By implementing this methodology, we aim to enhance the performance of models trained on the augmented dataset and mitigate potential biases and inaccuracies in fake news detection.

Dataset Augmentation Process for Fake News Classification:

- Obtaining Original Fake News Headlines:
 - A set of 1,299 original fake news headlines was acquired as the foundation for augmentation
- Preprocessing and Filtering:
 - The original headlines underwent preprocessing, filtering, and cleaning procedures.
 - Headlines exceeding a certain token length threshold were selected for further augmentation to ensure quality and relevance.
- Paraphrasing using a Pre-trained Bangla T5 Model:
 - A pre-trained Bangla T5 model designed for paraphrasing was employed.

- Each of the selected 1,190 headlines was used to generate four additional paraphrased versions using the Bangla T5 model.
- This process resulted in a total of 4,760 newly generated headlines, expanding the dataset significantly.
- Combining Original and Generated Headlines:
 - The total count of headlines reached 5,950 after combining the original and generated headlines.
 - To maintain the integrity of the original dataset, a careful recombination process was followed.
 - The original content, sources, labels, and dates were retained, while any duplicate paraphrased headlines were removed.
 - A total of 294 duplicated paraphrased data points were identified and eliminated.
 - The final augmented dataset consisted of 5,656 fake news headlines.

Handling Class Imbalance in the Dataset:

- Imbalanced Dataset:
 - The original dataset had an imbalance with 48,000 real news samples and 5,656 fake news samples.
- Addressing Class Imbalance:
 - To achieve a balanced dataset, an undersampling technique was applied to the larger class (real news).
 - 5,656 samples were randomly selected from the pool of real news samples.
 - The selected real news samples were combined with the existing 5,656 fake news samples.
- Resulting Balanced Dataset:
 - The balanced dataset contained a total of 11,312 samples, with 5,656 real news samples and 5,656 fake news samples.
 - This ensured an equal representation of both classes in the dataset.

Importance of Dataset Augmentation and Handling Class Imbalance:

- Dataset augmentation expands the dataset, providing a broader representation of fake news characteristics.
- It enhances the performance of models trained on the augmented dataset.
- Class imbalance techniques, such as undersampling or oversampling, are crucial to address the imbalanced nature of the dataset.
- These techniques help mitigate potential biases and inaccuracies in the model’s performance by ensuring equal representation of both classes.

TABLE 3.1: Models Comparison

Model	Dataset	Dataset Size	Data Used
AugFakeBERT	BanFakeNews-Full	47k (Imbalanced)	Headlines, Content
AugFakeBERT	BanFakeNews-Augmented/Balanced	8k (Balanced)	Headlines, Content
OurModel	BanFakeNews-Augmented	11k (Balanced)	Headlines
OurModel	BanFakeNews-Augmented	54k (Imbalanced)	Headlines

The creation of a balanced dataset was crucial for training and evaluating models in our case. With a balanced dataset, the models were exposed to an equal number of instances from both classes, promoting fair and unbiased learning. By providing an equal representation of real and fake news, the models better understood the characteristics and patterns associated with each class, leading to more accurate and reliable predictions.

The resulting balanced dataset of 11,312 samples was utilized for further analysis and evaluation.

We can clearly see that implementing our data augmentation process we were able to generate more data and curate the dataset into a more balanced dataset.

The architecture of the entire process is outlined in Figure 3.1

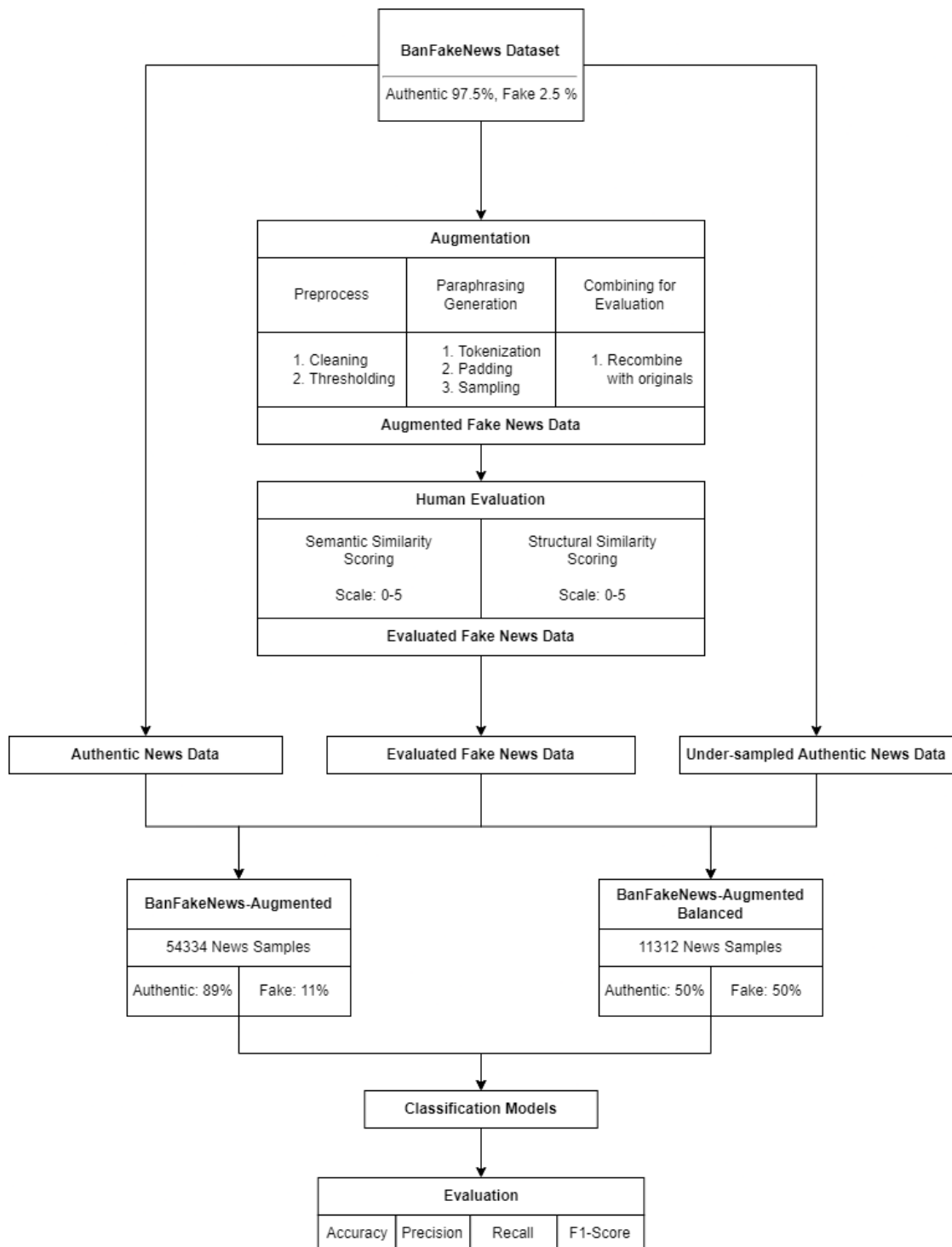


FIGURE 3.1: Architecture Overview

Chapter 4

Experimental Design

The experimental design section encompasses various aspects of the evaluation process and methodology used in the study. It includes the evaluation metrics employed, the dataset description, and the overall approach taken. The evaluation metrics focused on measuring semantic and structural similarity to assess the quality of the generated data. The dataset description highlights the BanFakeNews dataset used as the foundation and the incorporation of newly generated data to enhance it. The methodology involves data augmentation techniques, undersampling to address class imbalance, and the use of BanglaBERT for feature extraction. Parameters and settings specific to the classifiers' training and evaluation are also outlined.

4.1 Evaluation Metric

Evaluation of the generated data was conducted using a human evaluation approach. A subset of 200 random samples from the generated data was selected for evaluation, and two metrics, semantic similarity and structural similarity, were used to assess the quality of the generated samples.

The primary focus of the evaluation was to ensure a high level of semantic similarity between the generated samples and the original data. This metric aimed to measure the extent to which the generated samples captured the intended meaning and conveyed the same information as the original data.

Additionally, the evaluation aimed to achieve a low level of structural similarity between the generated samples and the original data. This metric assessed how different the

generated samples were in terms of their structure, syntax, and organization compared to the original data.

The evaluation results showed that the majority of the samples adhered to the desired metrics, as indicated by the survey average. This suggests that the generated samples successfully maintained a high level of semantic similarity while exhibiting low structural similarity to the original data.

TABLE 4.1: Dataset Evaluation Results

Evaluation Metric	Sample Count	Evaluators	Average Score
Semantic Similarity	200	5 people	4.08
Structural Similarity	200	5 people	2.40

4.2 Dataset Description

The dataset used in the experiments is referred to as the BanFakeNews dataset[1]. The BanFakeNews dataset served as the foundation for our research, and we further enhanced it by incorporating newly generated data. This combination of the original BanFakeNews dataset and the generated data resulted in an augmented and curated dataset.

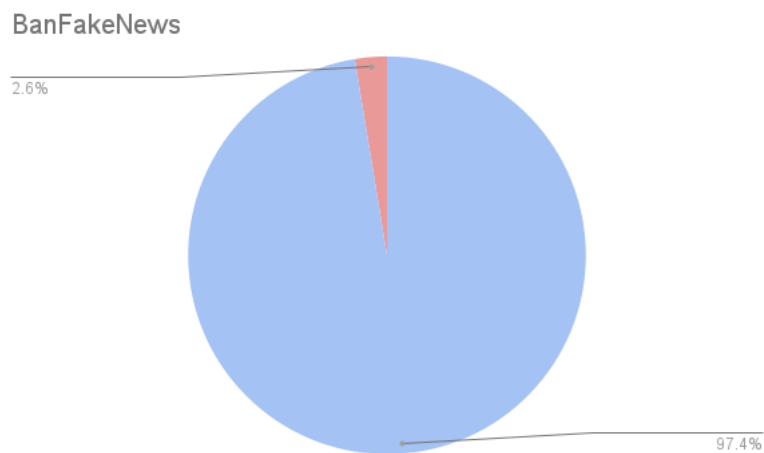


FIGURE 4.1: BanFakeNews Dataset Distribution

The original dataset used in our research consisted of 48,000 samples of authentic news and 1,000 samples of fake news. This dataset was initially imbalanced, with a significantly higher number of authentic news samples compared to fake news samples.

To address this class imbalance and enhance the dataset, we performed data augmentation techniques. By utilizing techniques such as paraphrasing and generating additional samples based on the existing data, we augmented the dataset with newly generated fake news samples. These generated samples aimed to increase the representation of fake news instances in the dataset, thereby improving the balance between the two classes.

After combining the augmented data with the original dataset, we obtained the imbalanced augmented dataset. The resulting dataset now comprised a larger number of fake news samples, while the number of authentic news samples remained the same. This combination aimed to create a more balanced representation of both classes and provide a more realistic training and evaluation environment for the classifiers.

By augmenting the original dataset and incorporating the generated fake news samples, we were able to address the initial class imbalance issue and create an imbalanced augmented dataset that contained a larger number of fake news samples than the original dataset. This dataset served as the basis for subsequent experiments and analyses, allowing for improved training and evaluation of the classifiers.

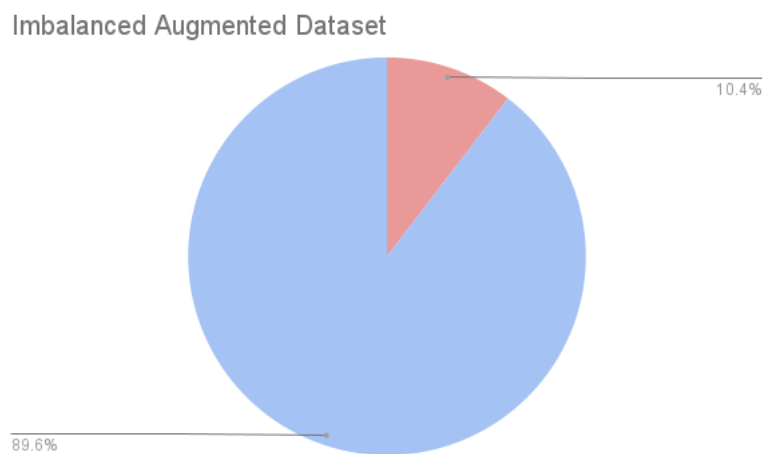


FIGURE 4.2: Imbalanced Augmented Dataset

Following the creation of the imbalanced augmented dataset, the next step in our approach was to address the class imbalance by performing undersampling[103]. Undersampling is a technique used to reduce the number of samples from the majority class (in this case, the authentic news class) to achieve a more balanced dataset.

In our case, the imbalanced augmented dataset initially contained a larger number of fake news samples and a smaller number of authentic news samples. To balance the dataset,

we applied undersampling specifically to the fake news class, reducing its sample count to match the number of authentic news samples.

By undersampling the fake news class, we aimed to create a balanced dataset that contained an equal number of samples for both classes. This balanced dataset then provides a more representative training environment for the classifiers, allowing them to learn from an equal proportion of authentic and fake news instances.

Undersampling helps prevent the classifiers from being biased towards the majority class, which in this case would be the authentic news class. It ensures that the models learn equally from both classes, leading to more accurate and reliable classification performance.

By performing undersampling on the imbalanced augmented dataset, we successfully achieved a balanced dataset with an equal number of authentic news and fake news samples. This balanced dataset served as the basis for subsequent training and evaluation of the classifiers, enabling them to make more informed and accurate predictions for both classes.

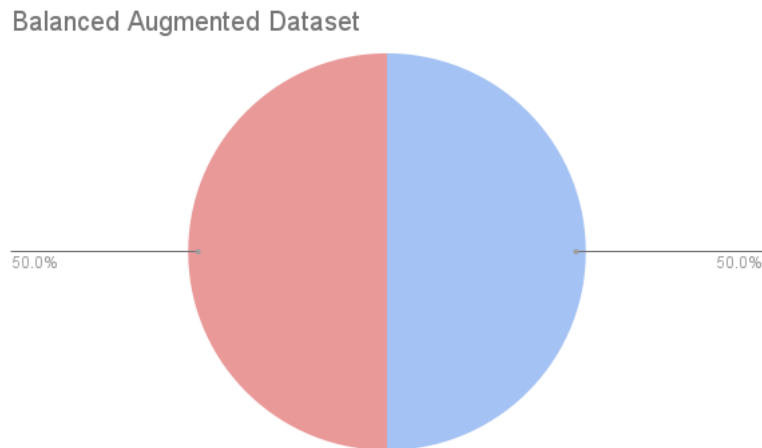


FIGURE 4.3: Balanced Augmented Dataset

TABLE 4.2: Dataset Distribution

Dataset	True News	Fake News	Total
BanFakeNews	48,678	1299	49,977
AugFake-BERT	4000	4000	8000
54k(50-50)	48678	5656	54334
11k (50:50)	5656	5656	11312

4.3 Experimental Setup

We performed classification on the 54k dataset and the 11k dataset. The classification was done using BanglaBert instead of Multilingual-Bert used in AugFakeBERT [2]

The training of classifiers involved several modifications and techniques to enhance their performance. Firstly, the dataset was divided into a 70-15-15 split for training, validation, and testing, respectively. This partitioning allowed for adequate training data while ensuring separate subsets for evaluation purposes.

To extract meaningful features from the Bangla language text, BanglaBERT was employed as the base model for feature extraction, replacing the use of Multilingual BERT. BanglaBERT is specifically designed for Bangla language processing and is expected to provide improved performance in capturing language-specific patterns and nuances. Please refer to the appropriate citations for more information on the BanglaBERT model.

To focus on the essential information within the dataset, only the headlines were used for training the classifiers. By utilizing the headline data, the models aimed to capture the key elements and patterns present in the shorter text, which are often vital for effective fake news detection.

To ensure efficient processing and maintain consistency, the sequence length for the headlines was set to 30 tokens. This involved truncating or padding the headlines to achieve a uniform length for training and inference.

TABLE 4.3: Parameters of Our Model

Hyperparameter	Value
Dropout Rate	0.1
Maximum Sequence Length	30
Number of Epochs	4
Batch Size	16
Activation Function	ReLU & Softmax
Learning Rate	2×10^{-3}
Optimizer	AdamW

A batch size of 16 was chosen for training the classifiers. This parameter determines the number of samples processed in each iteration during training. The selected batch size balanced the training efficiency and memory requirements.

TABLE 4.4: Parameters of AugFakeBERT

Hyperparameter	Value
Dropout Rate	None
Maximum Sequence Length	80
Number of Epochs	4
Batch Size	15
Activation Function	ReLU & Softmax
Learning Rate	2×10^{-3}
Optimizer	AdamW

To introduce randomness and improve generalization, a dropout layer with a dropout probability of 0.1 was incorporated into the model architecture. To account for potential variations due to randomness, the models were trained and evaluated four times, and the average performance across these scenarios was considered for a more robust evaluation.

While the aforementioned modifications were implemented, other hyperparameters, such as the learning rate, optimizer choice, loss function, and other relevant settings, remained consistent with the AugFakeBERT approach.

By applying these modifications and techniques, the classifiers were trained with the objective of enhancing their performance, addressing class imbalance, reducing training time, and effectively capturing the pertinent information from the headlines. The average performance across the four scenarios was considered to provide a reliable evaluation of the classifiers' capabilities.

TABLE 4.5: Experimental Setup

Dataset	M. Seq. Length	Learning Rate	Batch Size	Epochs	Class Weights	Dropout
54k (89:11)	30	2×10^{-3}	16	4	Yes	0.1
54k (89:11)	30	2×10^{-3}	16	4	No	0.1
11k (50:50)	30	2×10^{-3}	16	4	X	0.1

Chapter 5

Results and Discussions

5.1 Experimental Result

The experiments run on the above mentioned datasets with our model gives us the results as shown in figure 5.2 and 5.3.

Based on Table 5.1, the classification experiment results for the three datasets are presented. The F1 score, which is a measure of the model's overall performance, is highlighted. It is observed that the F1 score of the balanced augmented and curated dataset is higher, reaching 90.14%, compared to the imbalanced datasets.

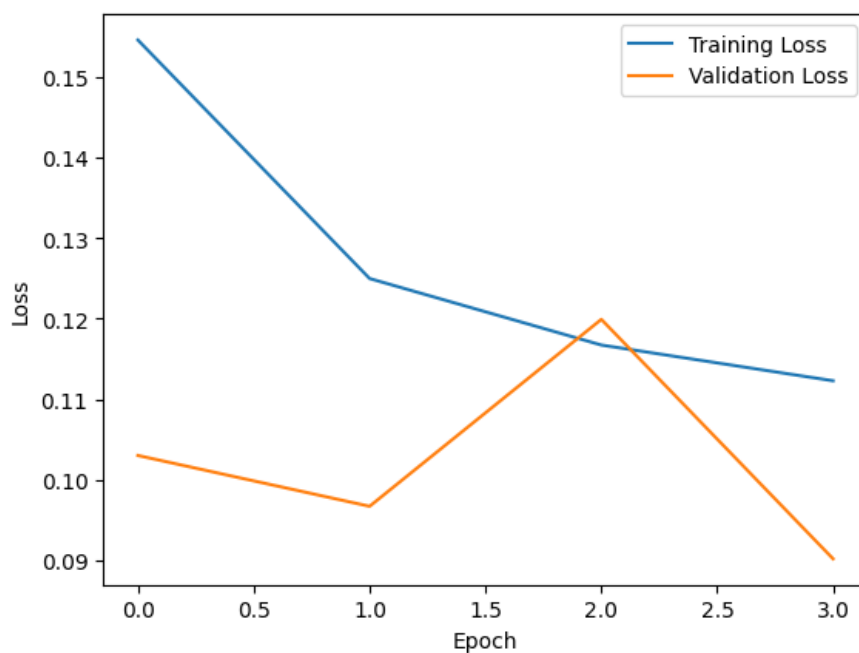


FIGURE 5.1: Imbalanced Dataset Loss with Dropout Layer but no class weights

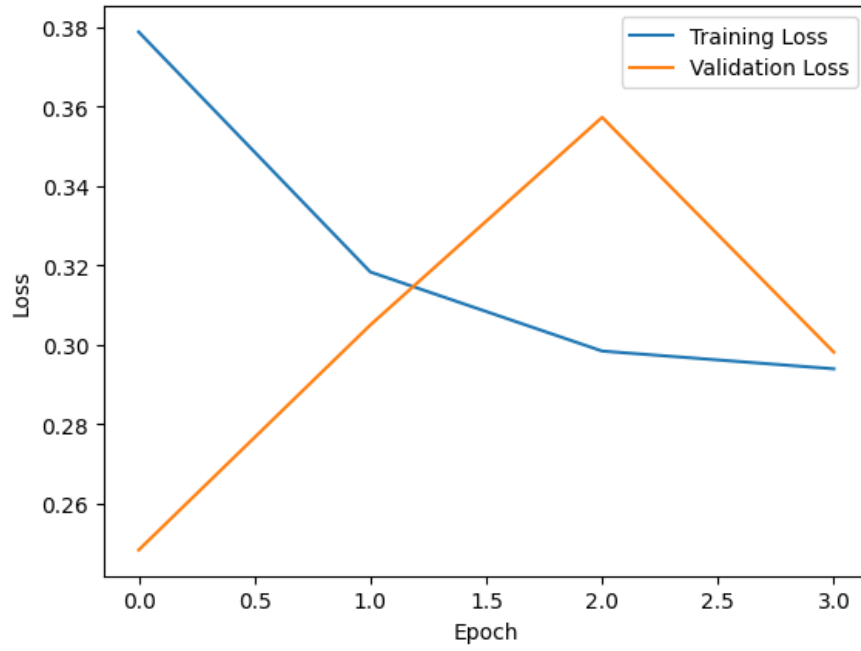


FIGURE 5.2: Imbalanced Dataset Loss with Dropout Layer and Class Weights

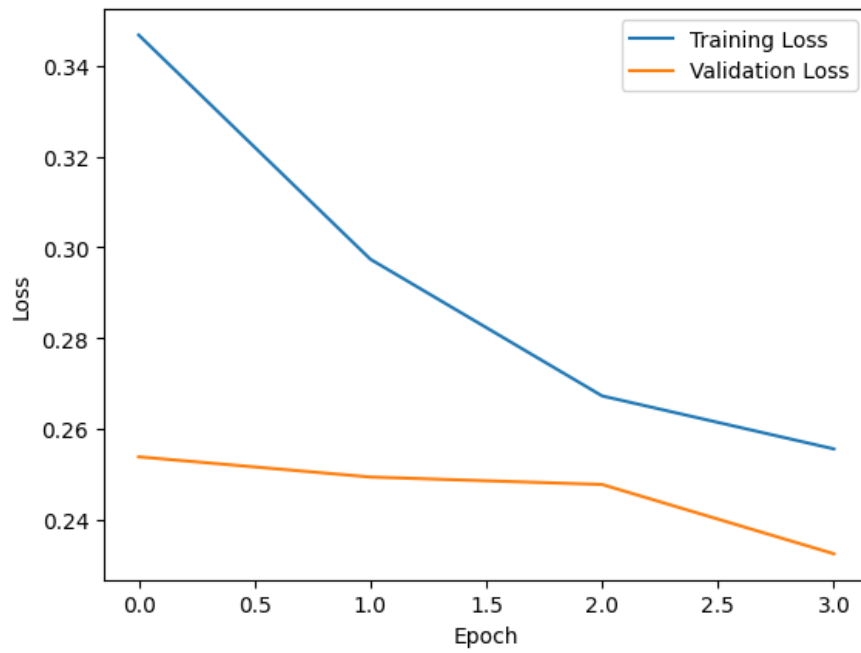


FIGURE 5.3: Balanced Dataset Loss with Dropout Layer

This finding suggests that the efforts made to balance the dataset through augmentation and curation have positively impacted the performance of the classifiers. By addressing the class imbalance issue and ensuring a more equal representation of real and fake news samples, the models have achieved better accuracy, precision, recall, or a combination thereof, resulting in a higher F1 score.

TABLE 5.1: Experimental Results

Dataset	A(%)	P(%)	R(%)	F1(%)
54k (89:11)	91.97	95.22	77.50	85.45
54k (89:11)	97.14	94.13	77.39	84.94
11k (50:50)	90.63	95.03	85.73	90.14

In the case of balancing the dataset, our model demonstrates superior precision, achieving a score of 95.03%. Moreover, it also achieves promising results, coming very close to state of the art in terms of accuracy and F1 score. This implies that our model effectively maintains high precision, correctly identifying fake news instances while maintaining a competitive overall accuracy and F1 score. This fact can be attributed to the change in dataset distribution.

By following the parameters of AugFakeBERT [2], we created a classifier model with hyper-parameters discussed in the table 4.3

We ran this model on the old imbalanced data set BanFakeNews [1], and also on our new combined data set as well as the newly created balanced data set.

TABLE 5.2: Results Comparison

Dataset	Imbalanced Dataset				Balanced Dataset			
	A (%)	P (%)	R (%)	F1 (%)	A (%)	P (%)	R (%)	F1 (%)
BanFakeNews [1]	97.80	48.70	50.00	49.34				-
BanFakeNews Augmented	91.97	95.22	77.50	85.45	90.63	95.03	85.73	90.14

The results in the table 5.2 are made clear with 2 key observations of the experimental results. We can see that, for the original balanced data set we have a very high accuracy value, but it is clear from the other validation scores that, the model tends to over fit to the data and is more biased towards the 'Authentic' label in the data set. This is because the overall bias in the data set is very large. So, the model tends to over fit and

learn this imbalance. In other words, it finds the lowest loss is just predicting 'Authentic' for any data point.

As the original data set did not include a balanced data set, we can see that there is no performance comparison. But, it is clear from the overall Accuracy, Recall, Precision and F1-Score that this balancing benefits the model with generalization and thus provides a more generalized model overall. Thus, the model doesn't have a very high accuracy, but it has a better recall and precision, which are both key elements of a generalized model.

In the "BanFakeNews" dataset, the model achieved a high accuracy of 97.80%. However, the precision was relatively low at 48.70%, indicating a significant number of false positives among the predicted positive instances. The recall score was 50.00%, meaning that only half of the actual positive instances were correctly identified. The F1-score, which considers both precision and recall, was 49.34

On the other hand, in the "BanFakeNews-Augmented" dataset, the model achieved a slightly lower accuracy of 91.97%. However, there was a substantial improvement in precision, which increased to 95.22%. This suggests that a higher proportion of true positives were identified among the predicted positive instances. The recall score also improved to 77.50%, indicating a higher percentage of actual positive instances captured by the model. As a result, the F1-score showed a significant enhancement, reaching 85.45

In summary, the "BanFakeNews-Augmented" dataset demonstrated better performance metrics compared to the original "BanFakeNews" dataset. The augmented dataset allowed the model to achieve higher precision, recall, and F1-score. This improvement suggests that the augmentation process likely helped the model better handle the imbalanced nature of the dataset and improve its ability to correctly identify positive instances.

In the imbalanced version of the "BanFakeNews-Augmented" dataset, the model achieved a relatively high accuracy of 90.63%. However, the precision was slightly lower at 95.03%, indicating a proportion of false positives among the predicted positive instances. The recall score was 85.73%, suggesting that the model captured a substantial percentage of the actual positive instances. The F1-score, combining precision and recall, was 90.14%.

In contrast, the balanced version of the "BanFakeNews-Augmented" dataset demonstrated slightly lower accuracy at 91.97%. However, there was a noticeable improvement in precision, which increased to 95.22%. This indicates a higher proportion of true positives among the predicted positive instances. The recall score also improved to 85.73%, suggesting that the model continued to identify a significant percentage of the actual positive instances. Consequently, the F1-score showed an enhancement, reaching 90.14%.

Overall, the performance difference between the imbalanced and balanced versions of the "BanFakeNews-Augmented" dataset is relatively subtle. While the balanced dataset exhibited slightly better precision, the recall and F1-score remained the same. This suggests that the augmentation process played a more significant role in improving the model's ability to handle the imbalanced class distribution, rather than the balancing technique itself.

It's worth noting that even though the balanced dataset did not significantly outperform the imbalanced dataset, achieving high precision is valuable, particularly in applications where false positives can have severe consequences. Therefore, using a balanced dataset, augmented or not, can be beneficial in improving the model's performance and reliability.

5.2 Limitations

Although the augmented data shows higher semantic similarity and lower structure relevance, however some of the limitations of this research are:

- Preserving semantic similarity during data augmentation is complex, and there may be cases where the augmented texts do not accurately capture the intended meaning or context of the original texts[92].
- The effectiveness of data augmentation techniques in generating high-quality and diverse augmented fake news texts is crucial for the success of the research. But, it is essential to evaluate the generated augmented data to ensure that it effectively represents the variations and patterns present in real-world fake news.
- While the research aims to construct a balanced dataset by augmenting the existing data, there is a risk of overfitting the models to the augmented dataset [104]. This can limit the generalization of the models to real-world scenarios where the distribution of fake news may differ.
- The choice of evaluation metrics for assessing the accuracy of the machine learning and deep learning models is not generalized. Therefore, potential biases in the dataset or evaluation process should be identified and addressed to ensure unbiased evaluations.

Chapter 6

Conclusion and Future Work

In our study, we acknowledge certain limitations and identify potential areas for future work. Firstly, due to resource and time constraints, we focused solely on utilizing headlines of the BanFakeNews dataset [1] for our experiments. While headlines provide valuable information, augmenting the content data, such as the body of the news articles, could offer further improvements in addressing data imbalance and achieving higher accuracy.

By extending our augmentation framework to include content data, we could potentially generate synthetic examples that align with the original data distribution. This augmentation approach can help address the issue of data imbalance, where the number of samples in different classes is significantly different [2]. By augmenting the content data, we can create a more balanced dataset, which can lead to improved performance and accuracy of models used for fake news detection or related tasks.

In future work, it would be beneficial to explore the augmentation of content data using techniques such as paraphrasing, backtranslation, or other language generation methods [73]. This expanded approach would allow for a more comprehensive analysis of both headlines and content, capturing the nuances and subtleties present in the full text of news articles. Additionally, further research could be conducted to investigate the impact of augmented content data on the performance of machine learning models. This could involve evaluating the accuracy, precision, recall, and other metrics to quantify the effectiveness of the augmented dataset in reducing bias and enhancing the overall performance of the models.

By addressing these limitations and exploring the augmentation of both headlines and content data, we can potentially overcome data constraints and achieve even better results in fake news detection and related tasks.

Bibliography

- [1] M. Z. Hossain, M. A. Rahman, M. S. Islam, and S. Kar, “Banfakenews: A dataset for detecting fake news in bangla,” *arXiv preprint arXiv:2004.08789*, 2020.
- [2] A. J. Keya, M. A. H. Wadud, M. Mridha, M. Alatiyyah, and M. A. Hamid, “Augfake-bert: Handling imbalance through augmentation of fake news using bert to enhance the performance of fake news classification,” *Applied Sciences*, vol. 12, no. 17, p. 8398, 2022.
- [3] O. F. Rakib, S. Akter, M. A. Khan, A. K. Das, and K. M. Habibullah, “Bangla word prediction and sentence completion using gru: an extended version of rnn on n-gram language model,” in *2019 International Conference on Sustainable Technologies for Industry 4.0 (STI)*. IEEE, 2019, pp. 1–6.
- [4] Y. Kang, Z. Cai, C.-W. Tan, Q. Huang, and H. Liu, “Natural language processing (nlp) in management research: A literature review,” *Journal of Management Analytics*, vol. 7, no. 2, pp. 139–172, 2020.
- [5] J. J. Webster and C. Kit, “Tokenization as the initial phase in nlp,” in *COLING 1992 volume 4: The 14th international conference on computational linguistics*, 1992.
- [6] K. Chowdhary and K. Chowdhary, “Natural language processing,” *Fundamentals of artificial intelligence*, pp. 603–649, 2020.
- [7] R. Socher, Y. Bengio, and C. D. Manning, “Deep learning for nlp (without magic),” in *Tutorial Abstracts of ACL 2012*, 2012, pp. 5–5.
- [8] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.
- [9] X. Zhang and A. A. Ghorbani, “An overview of online fake news: Characterization, detection, and discussion,” *Information Processing & Management*, vol. 57, no. 2, p. 102025, 2020.

- [10] J. C. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, "Supervised learning for fake news detection," *IEEE Intelligent Systems*, vol. 34, no. 2, pp. 76–81, 2019.
- [11] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–40, 2020.
- [12] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text data augmentation for deep learning," *Journal of big Data*, vol. 8, pp. 1–34, 2021.
- [13] J. S. Liu and Y. N. Wu, "Parameter expansion for data augmentation," *Journal of the American Statistical Association*, vol. 94, no. 448, pp. 1264–1274, 1999.
- [14] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for nlp," *arXiv preprint arXiv:2105.03075*, 2021.
- [15] S. Li, M. Xie, K. Gong, C. H. Liu, Y. Wang, and W. Li, "Transferable semantic augmentation for domain adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 516–11 525.
- [16] Y. Wang, X. Pan, S. Song, H. Zhang, G. Huang, and C. Wu, "Implicit semantic data augmentation for deep networks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [17] Y. Nie, Y. Tian, X. Wan, Y. Song, and B. Dai, "Named entity recognition for social media texts with semantic augmentation," *arXiv preprint arXiv:2010.15458*, 2020.
- [18] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell, "Understanding data augmentation for classification: when to warp?" in *2016 international conference on digital image computing: techniques and applications (DICTA)*. IEEE, 2016, pp. 1–6.
- [19] S. Kotsiantis, D. Kanellopoulos, P. Pintelas *et al.*, "Handling imbalanced datasets: A review," *GESTS international transactions on computer science and engineering*, vol. 30, no. 1, pp. 25–36, 2006.
- [20] M. S. Rahman, F. B. Ashraf, and M. R. Kabir, "An efficient deep learning technique for bangla fake news detection," in *2022 25th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2022, pp. 206–211.
- [21] D. Ramyachitra and P. Manikandan, "Imbalanced dataset classification and solutions: a review," *International Journal of Computing and Business Research (IJCBR)*, vol. 5, no. 4, pp. 1–29, 2014.

- [22] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Machine learning*, vol. 15, pp. 201–221, 1994.
- [23] Y. Wang, G. Huang, S. Song, X. Pan, Y. Xia, and C. Wu, "Regularizing deep networks with semantic data augmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 7, pp. 3733–3748, 2021.
- [24] S. B. S. Mugdha, S. M. Ferdous, and A. Fahmin, "Evaluating machine learning algorithms for bengali fake news detection," in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2020, pp. 1–6.
- [25] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.
- [26] A. Magueresse, V. Carles, and E. Heetderks, "Low-resource languages: A review of past work and future challenges," *arXiv preprint arXiv:2006.07264*, 2020.
- [27] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," *Language and cognitive processes*, vol. 6, no. 1, pp. 1–28, 1991.
- [28] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the 2011 conference on empirical methods in natural language processing*, 2011, pp. 262–272.
- [29] M. Q. Patton *et al.*, "Qualitative evaluation methods," 1980.
- [30] G. Lee, I. Lee, H. Ha, K.-G. Lee, H. Hyun, A. Shin, and B.-G. Chun, "Refurbish your training data: Reusing partially augmented samples for faster deep neural network training." in *USENIX Annual Technical Conference*, 2021, pp. 537–550.
- [31] M. C. Ramos, "Some ethical implications of qualitative research," *Research in Nursing & Health*, vol. 12, no. 1, pp. 57–63, 1989.
- [32] T. S. Apon, R. Anan, E. A. Modhu, A. Suter, I. J. Sneha, and M. G. R. Alam, "Banglasarc: A dataset for sarcasm detection," in *2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. IEEE, 2022, pp. 1–5.
- [33] M. E. Markiewicz and C. J. de Lucena, "Object oriented framework development," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 7, no. 4, pp. 3–9, 2001.
- [34] L. Pevzner and M. A. Hearst, "A critique and improvement of an evaluation metric for text segmentation," *Computational Linguistics*, vol. 28, no. 1, pp. 19–36, 2002.

- [35] I. Hernández, S. Sawicki, F. Roos-Frantz, and R. Z. Frantz, “Cloud configuration modelling: a literature review from an application integration deployment perspective,” *Procedia Computer Science*, vol. 64, pp. 977–983, 2015.
- [36] S. Frühwirth-Schnatter, “Data augmentation and dynamic linear models,” *Journal of time series analysis*, vol. 15, no. 2, pp. 183–202, 1994.
- [37] A. Antoniou, A. Storkey, and H. Edwards, “Data augmentation generative adversarial networks,” *arXiv preprint arXiv:1711.04340*, 2017.
- [38] C. Shorten and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.
- [39] J. Wei and K. Zou, “Eda: Easy data augmentation techniques for boosting performance on text classification tasks,” *arXiv preprint arXiv:1901.11196*, 2019.
- [40] R. Keskisärkkä, “Automatic text simplification via synonym replacement,” 2012.
- [41] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond empirical risk minimization,” *arXiv preprint arXiv:1710.09412*, 2017.
- [42] L. Zhang, Z. Deng, K. Kawaguchi, A. Ghorbani, and J. Zou, “How does mixup help with robustness and generalization?” *arXiv preprint arXiv:2010.04819*, 2020.
- [43] E. Alemayehu and Y. Fang, “A submodular optimization framework for imbalanced text classification with data augmentation,” *IEEE Access*, 2023.
- [44] J. Chen, Z. Yang, and D. Yang, “Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification,” *arXiv preprint arXiv:2004.12239*, 2020.
- [45] Y. Wang, J. Yan, Z. Yang, Y. Zhao, and T. Liu, “Optimizing gis partial discharge pattern recognition in the ubiquitous power internet of things context: A mixnet deep learning model,” *International Journal of Electrical Power & Energy Systems*, vol. 125, p. 106484, 2021.
- [46] R. Sennrich, B. Haddow, and A. Birch, “Improving neural machine translation models with monolingual data,” *arXiv preprint arXiv:1511.06709*, 2015.
- [47] T. Stoev, A. Ferrario, B. Demiray, M. Luo, M. Martin, and K. Yordanova, “Coping with imbalanced data in the automated detection of reminiscence from everyday life conversations of older adults,” *IEEE Access*, vol. 9, pp. 116 540–116 551, 2021.
- [48] S. Kobayashi, “Contextual augmentation: Data augmentation by words with paradigmatic relations,” *arXiv preprint arXiv:1805.06201*, 2018.

- [49] Y. Yang, C. Malaviya, J. Fernandez, S. Swayamdipta, R. L. Bras, J.-P. Wang, C. Bhagavatula, Y. Choi, and D. Downey, “Generative data augmentation for commonsense reasoning,” *arXiv preprint arXiv:2004.11546*, 2020.
- [50] S. Vosoughi, D. Roy, and S. Aral, “The spread of true and false news online,” *science*, vol. 359, no. 6380, pp. 1146–1151, 2018.
- [51] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, “Automatic detection of fake news,” *arXiv preprint arXiv:1708.07104*, 2017.
- [52] Y. Long, Q. Lu, R. Xiang, M. Li, and C.-R. Huang, “Fake news detection through multi-perspective speaker profiles,” in *Proceedings of the eighth international joint conference on natural language processing (volume 2: Short papers)*, 2017, pp. 252–256.
- [53] F. Yang, A. Mukherjee, and E. Dragut, “Satirical news detection and analysis using attention mechanism and linguistic features,” *arXiv preprint arXiv:1709.01189*, 2017.
- [54] G. Karadzhov, P. Nakov, L. Màrquez, A. Barrón-Cedeño, and I. Koychev, “Fully automated fact checking using external sources,” *arXiv preprint arXiv:1710.00341*, 2017.
- [55] X. Dong, U. Victor, S. Chowdhury, and L. Qian, “Deep two-path semi-supervised learning for fake news detection,” *arXiv preprint arXiv:1906.05659*, 2019.
- [56] P. S. Ray *et al.*, “Bengali language handbook.” 1966.
- [57] M. Lan, Z. Zhang, Y. Lu, and J. Wu, “Three convolutional neural network-based models for learning sentiment word vectors towards sentiment analysis,” in *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016, pp. 3172–3179.
- [58] H. Saleh, A. Alharbi, and S. H. Alsamhi, “Opcnn-fake: optimized convolutional neural network for fake news detection,” *IEEE Access*, vol. 9, pp. 129 471–129 489, 2021.
- [59] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi, and B.-W. On, “Fake news stance detection using deep learning architecture (cnn-lstm),” *IEEE Access*, vol. 8, pp. 156 695–156 706, 2020.
- [60] O. Ajao, D. Bhowmik, and S. Zargari, “Fake news identification on twitter with hybrid cnn and rnn models,” in *Proceedings of the 9th international conference on social media and society*, 2018, pp. 226–230.

- [61] S. Singhanian, N. Fernandez, and S. Rao, “3han: A deep neural network for fake news detection,” in *International conference on neural information processing*. Springer, 2017, pp. 572–581.
- [62] N. Alosbhan, “Act: Automatic fake news classification through self-attention,” in *12th ACM Conference on Web Science*, 2020, pp. 115–124.
- [63] Y.-J. Lu and C.-T. Li, “Gcan: Graph-aware co-attention networks for explainable fake news detection on social media,” *arXiv preprint arXiv:2004.11648*, 2020.
- [64] H. Jwa, D. Oh, K. Park, J. M. Kang, and H. Lim, “exbake: Automatic fake news detection model based on bidirectional encoder representations from transformers (bert),” *Applied Sciences*, vol. 9, no. 19, p. 4062, 2019.
- [65] T. Zhang, D. Wang, H. Chen, Z. Zeng, W. Guo, C. Miao, and L. Cui, “Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection,” in *2020 international joint conference on neural networks (IJCNN)*. IEEE, 2020, pp. 1–8.
- [66] R. K. Kaliyar, A. Goswami, and P. Narang, “Fakebert: Fake news detection in social media with a bert-based deep learning approach,” *Multimedia tools and applications*, vol. 80, no. 8, pp. 11 765–11 788, 2021.
- [67] C.-L. Wu, H.-P. Hsieh, J. Jiang, Y.-C. Yang, C. Shei, and Y.-W. Chen, “Muffle: Multi-modal fake news influence estimator on twitter,” *Applied Sciences*, vol. 12, no. 1, p. 453, 2022.
- [68] S. Hiriyannaiah, A. Srinivas, G. K. Shetty, G. Siddesh, and K. Srinivasa, “A computationally intelligent agent for detecting fake news using generative adversarial networks,” in *Hybrid Computational Intelligence*. Elsevier, 2020, pp. 69–96.
- [69] N. J. Ria, S. A. Khushbu, M. A. Yousuf, A. K. M. Masum, S. Abujar, and S. A. Hossain, “Toward an enhanced bengali text classification using saint and common form,” in *2020 11th international conference on computing, communication and networking technologies (ICCCNT)*. IEEE, 2020, pp. 1–5.
- [70] K. Knight and J. Graehl, “Machine transliteration,” *arXiv preprint cmp-lg/9704003*, 1997.
- [71] M. E. Peters, M. Neumann, R. L. Logan IV, R. Schwartz, V. Joshi, S. Singh, and N. A. Smith, “Knowledge enhanced contextual word representations,” *arXiv preprint arXiv:1909.04164*, 2019.

- [72] L. M. Rose, N. Matragkas, D. S. Kolovos, and R. F. Paige, "A feature model for model-to-text transformation languages," in *2012 4th International Workshop on Modeling in Software Engineering (MISE)*. IEEE, 2012, pp. 57–63.
- [73] S. Edunov, M. Ott, M. Auli, and D. Grangier, "Understanding back-translation at scale," *arXiv preprint arXiv:1808.09381*, 2018.
- [74] C. A. Ferguson and M. Chowdhury, "The phonemes of bengali," *Language*, vol. 36, no. 1, pp. 22–59, 1960.
- [75] B. Hayes and A. Lahiri, "Bengali intonational phonology," *Natural language & linguistic theory*, vol. 9, pp. 47–96, 1991.
- [76] R. R. Chowdhury, M. S. Hossain, R. ul Islam, K. Andersson, and S. Hossain, "Bangla handwritten character recognition using convolutional neural network with data augmentation," in *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. IEEE, 2019, pp. 318–323.
- [77] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [78] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, "Masked language model scoring," *arXiv preprint arXiv:1910.14659*, 2019.
- [79] Y. Sun, Y. Zheng, C. Hao, and H. Qiu, "Nsp-bert: A prompt-based zero-shot learner through an original pre-training task—next sentence prediction," *arXiv preprint arXiv:2109.03564*, 2021.
- [80] K. M. Hosny, M. A. Kassem, and M. M. Foad, "Classification of skin lesions using transfer learning and augmentation with alex-net," *PloS one*, vol. 14, no. 5, p. e0217293, 2019.
- [81] A. Karadeniz, "Cohesion and coherence in written texts of students of faculty of education." *Journal of Education and Training Studies*, vol. 5, no. 2, pp. 93–99, 2017.
- [82] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555*, 2020.
- [83] Z. Chi, S. Huang, L. Dong, S. Ma, B. Zheng, S. Singhal, P. Bajaj, X. Song, X.-L. Mao, H. Huang *et al.*, "Xlm-e: Cross-lingual language model pre-training via electra," *arXiv preprint arXiv:2106.16138*, 2021.

- [84] S. Ni and H.-Y. Kao, “Electra is a zero-shot learner, too,” *arXiv preprint arXiv:2207.08141*, 2022.
- [85] A. Bhattacharjee, T. Hasan, W. U. Ahmad, K. Samin, M. S. Islam, A. Iqbal, M. S. Rahman, and R. Shahriyar, “Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla,” *arXiv preprint arXiv:2101.00204*, 2021.
- [86] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [87] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” *Advances in neural information processing systems*, vol. 32, 2019.
- [88] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using rnn encoder-decoder for statistical machine translation,” *arXiv preprint arXiv:1406.1078*, 2014.
- [89] W. Yin, K. Kann, M. Yu, and H. Schütze, “Comparative study of cnn and rnn for natural language processing,” *arXiv preprint arXiv:1702.01923*, 2017.
- [90] A. Sherstinsky, “Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network,” *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.
- [91] I. Rish *et al.*, “An empirical study of the naive bayes classifier,” in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.
- [92] J. Kandola, N. Cristianini, and J. Shawe-taylor, “Learning semantic similarity,” *Advances in neural information processing systems*, vol. 15, 2002.
- [93] A. Das and D. Saha, “Deep learning based bengali question answering system using semantic textual similarity,” *Multimedia Tools and Applications*, pp. 1–25, 2022.
- [94] A. S. Bauer, P. Schmaus, F. Stulp, and D. Leidner, “Probabilistic effect prediction through semantic augmentation and physical simulation,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9278–9284.

- [95] J. Li, X. Zhang, X. Zhou *et al.*, “Albert-based self-ensemble model with semisupervised learning and data augmentation for clinical semantic textual similarity calculation: algorithm validation study,” *JMIR Medical Informatics*, vol. 9, no. 1, p. e23086, 2021.
- [96] G. Szlobodnyik and L. Farkas, “Data augmentation by guided deep interpolation,” *Applied Soft Computing*, vol. 111, p. 107680, 2021.
- [97] M. A. Iqbal, O. Sharif, M. M. Hoque, and I. H. Sarker, “Word embedding based textual semantic similarity measure in bengali,” *Procedia Computer Science*, vol. 193, pp. 92–101, 2021.
- [98] M. Shajalal and M. Aono, “Semantic textual similarity in bengali text,” in *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*. IEEE, 2018, pp. 1–5.
- [99] A. Sarkar and M. S. Hossen, “Automatic bangla text summarization using term frequency and semantic similarity approach,” in *2018 21st International Conference of Computer and Information Technology (ICCIT)*. IEEE, 2018, pp. 1–6.
- [100] A. Akil, N. Sultana, A. Bhattacharjee, and R. Shahriyar, “Banglaparaphrase: A high-quality bangla paraphrase dataset,” *arXiv preprint arXiv:2210.05109*, 2022.
- [101] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, “mt5: A massively multilingual pre-trained text-to-text transformer,” *arXiv preprint arXiv:2010.11934*, 2020.
- [102] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.
- [103] X.-Y. Liu, J. Wu, and Z.-H. Zhou, “Exploratory undersampling for class-imbalance learning,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 2, pp. 539–550, 2008.
- [104] L. Rice, E. Wong, and Z. Kolter, “Overfitting in adversarially robust deep learning,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 8093–8104.