



**Creating Synthetic Data for Stance Detection Tasks using Large Language
Models**

Author: Alahnouf Saud S Alsemairi

Supervisor: Dr. Fernando Alva Manchego

A dissertation submitted in partial fulfilment of the requirements for the degree
of:

MSc Artificial Intelligence

School of Computer Science and Informatics

Cardiff University

September 2023

Acknowledgement

At the completion of this study, I would like to thank Allah, who empowered me throughout this dissertation. Words alone cannot express my thanks.

I extend my appreciation and gratitude to my supervisor, Dr. Fernando Alva Manchego, who patiently guided and encouraged me during this dissertation. Your words of encouragement have been valuable to me.

I would also like to thank my family, who played an important role in supporting and believing in me throughout my journey. Without them, it would not have been possible.

Abstract

Stance detection is a natural language processing (NLP) task that analyses people's stances (e.g. in favour, against or neutral) towards a specific topic. It is usually tackled using supervised classification approaches. However, collecting datasets with suitable human annotations is a resource-expensive process. The impressive capability of large language models (LLMs) in generating human-like text has revolutionized various NLP tasks. Therefore, in this dissertation, we investigate the capabilities of LLMs, specifically ChatGPT and Falcon, as a potential solution to create synthetic data that may address the data scarcity problem in stance detection tasks, and observe its impact on the performance of stance detection models. The study was conducted across various topics (e.g. Feminism, Covid-19) and two languages (English and Arabic). Different prompting approaches were employed to guide these LLMs in generating artificial data that is similar to real-world data. The results demonstrate a range of capabilities and limitations of LLMs for this use case. ChatGPT's ethical guidelines affect its performance in simulating real-world tweets. Conversely, the open-source Falcon model's performance in resembling the original data was better than ChatGPT's; however, it could not create good Arabic tweets compared to ChatGPT. The study concludes that the current abilities of ChatGPT and Falcon are insufficient to generate diverse synthetic tweets. Thus, additional improvements are required to bridge the gap between synthesized and real-world data to enhance the performance of stance detection models.

Table of Contents

Chapter1: Introduction	1
1.1 Research Motivation	1
1.2 Research statement.....	2
1.3 Contributions.....	2
1.4 Research Aim and Objectives	2
1.5 Road Map.....	3
Chapter 2: Literature Review	4
2.1 Overview	4
2.2 Background	4
2.2.1 Stance detection.....	4
2.2.2 Transformers.....	6
2.2.3 Large language models.....	6
2.2.3.1 ChatGPT	7
2.2.3.2 Falcon:	8
2.1.4 Synthetic Dataset.....	9
2.3 Related Works	10
2.3.1 Automatic stance detection	10
2.3.2 Data augmentation using LLMs	11
2.3.3 Annotation process using LLMs	13
2.3.4 Prompt-based learning:	14
2.4 Summary	16
Chapter 3: Methodology.....	17
3.1 Datasets and models preparation	17
3.2 Evaluation	18
Chapter 4: Implementation.....	21
4.1 Overview	21
4.2 Dataset descriptions	21
4.3 Prompt strategies.....	23
4.3.1 Prompt for ChatGPT	23
4.3.1.1 General and simple prompts for all stance	23
4.3.1.2 Specific prompt for each stance.....	25
4.3.1.3 Template prompt for each stance	27

4.3.1.4 Arabic tweet prompts:.....	28
4.3.2 Prompts for the Falcon Model:.....	29
4.3.3 Brief description for the generated data:	30
4.3.4 Topic modelling.....	30
4.4 Data Pre-processing	32
4.5 Model selection	33
Chapter 5: Results and Evaluation	34
5.1 Overview	34
5.2 Experiments with Feminism Movements datasets.....	34
5.3 Experiments with Covid-19 Vaccine datasets.....	38
5.4 Experiments with Women’s Empowerment datasets in Arabic	41
5.5 Main Findings	43
Chapter 6: Conclusion	46
1.1 Conclusion	46
2.2 Limitations	47
2.3 Future works	47
2.4 Reflection.....	48
References	49
Appendix A	56
Appendix B	60

List of Figures

Figure 1. A prompt by Zhang et al. (2023)	15
Figure 2. Overall approach with each of the LLMs.....	20
Figure 3. General and simple prompt for all stances	24
Figure 4. A sample of generated tweets using General and simple prompts for all Stance approach.	24
Figure 5. Specific prompt for each stance	26
Figure 6. Template base prompt for each stance.....	27
Figure 7. Simple prompt for the Falcon 7B-instruct model.....	30
Figure 8. Confusion matrix for RoBERTa (Feminism)	37
Figure 9. Confusion matrix for RoBERTa (Covid-19)	40
Figure 10. Confusion matrix for ArabBERT.....	43
Figure 11. ChatGPT responses towards offensive content requests.	44

List of Tables

Table 1. A summary of the datasets	23
Table 2. Inferred topics for each stance from datasets in English and Arabic	32
Table 3. Testing results for training ChatGPT's data for Femininsm topic	35
Table 4. F1 scores for each stance. (ChatGPT- Feminism).....	36
Table 5. Testing results for training Falcon's data for Femininsm topic.....	37
Table 6. Samples of tweets with offensive and negative tones	38
Table 7. Testing results for training ChatGPT's data for Covid-19 topic	39
Table 8. F1 scores for each stance. (ChatGPT- Covid-19).....	40
Table 9. Testing results for training Falcon's data for Covid-19 topic.....	40
Table 10 Testing results for training ChatGPT's data for Women's Empowerment topic	42
Table 11. F1 scores for each stance. (ChatGPT- Women's Empowerment)	42
Table 12. Samples of each dataset show the diversity issue	45

List of Abbreviations

LLM	Large language model
NLP	Natural Language processing
API	Application Programming Interface
GPT	Generative Pre-trained Transformer
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Pretraining Approach
MLM	Masked Language Modelling
RLHF	Reinforcement Learning from Human Feedback
CoLA	Corpus of Linguistic Acceptability
TII	Technology Innovation Institute
SVM	Support Vector Machine
MTurk	Amazon Mechanical Turk
CoT	Chain of Thought

Chapter1: Introduction

1.1 Research Motivation

Stance detection consists of determining whether an author's perspective or stance towards a specific topic or domain is oppositional, supportive, or neutral (Küçük and Can 2020). It thus plays a crucial role in the analysis of peoples' opinions across various social media platforms such as X (formerly known as Twitter) and Facebook. Traditional approaches for automatic stance detection involve training models in a supervised fashion, where annotated data is a vital requirement (Alturayef et al. 2023). However, obtaining labelled datasets for stance detection with respect to specific targets, such as feminism or the BREXIT referendum, can be challenging and time-consuming.

Recent natural language processing (NLP) advancements in large language models (LLMs) like ChatGPT have demonstrated that these models have remarkable capabilities in natural language understanding and text generation. They can thus be utilised across a diverse range of tasks, including text summarisation, question-answering, code generation, and the solving of mathematical equations (Liu et al. 2023; Alexandr et al. 2021).

Generating synthetic datasets using large language models offers one way for researchers to overcome the limitations posed by the availability of annotated data, especially for recent or specific domains. Møller et al. (2023) and Veselovsky et al. (2023) recently demonstrated the ability of such models to produce synthetic data that adequately resembles real-world texts by rewriting existing texts or by writing similar texts to those in existence, offering an effective and scalable solution to the data scarcity problem.

The motivation behind this study is the exploration of the benefits and possibilities of leveraging large language models to create new synthetic datasets to be applied to stance detection classification problems across specific topics. By utilising the generative power of LLMs, the aim is thus to create a wider range of text samples that are contextually appropriate as well as reflective of a variety of different stances and viewpoints toward the target topic.

1.2 Research statement

The limited sizes of existing annotated datasets as developed by human annotators commonly affect the learning process of the fine-tuned models for stance detection tasks. This limitation can impact the model's ability to effectively generalise to unseen data. To overcome this issue, new synthetic data can be created using LLMs as a way to increase the volume of training datasets for use in stance detection models, hence improving these models' performance.

In this study, we intensively investigate the performance of LLMs, specifically ChatGPT and Falcon, as potential solutions to acquire labelled datasets. The investigation covers the development of dataset for two different topics (the Feminist Movement and Covid-19 Vaccines), and two languages (English and Arabic), as well as adopting various fine-tuned models for stance detection tasks. The aim was to assess LLMs' ability to mimic real-world training sets contextually. To achieve this, various prompt approaches were designed to guide the LLMs in the data generation process. Through these prompts, we intended to capture various perspectives and tones in the original data to enhance the model's performance in classifying stances.

1.3 Contributions

This study contributes to understanding the capabilities of LLMs in creating new datasets for stance detection task. In particular, it provides valuable insights into ChatGPT's style in data generation across two topics. Through extensive experimentation, it demonstrates how the ethical guidelines and constraints imposed by OpenAI influenced ChatGPT performance in terms of simulating real-world characteristics. It also explores the abilities of open-source LLMs, particularly Falcon, in a similar manner. Such exploration includes identifying any limitations in the data creation process. Notably, we contribute to having insights into the bilingual capabilities of Falcon and ChatGPT by using them to create Arabic data, allowing a better understanding of their performance beyond the English language.

1.4 Research Aim and Objectives

The study aimed to examine the performance of ChatGPT and Falcon to resemble human comments toward specific targets. It then sought to analyse the performance of models

trained using both the developed synthetic and existing actual datasets to check whether the use of the former can enhance stance detection model performances. The objectives, therefore, can be summarised as follows:

- Design appropriate prompts to guide the selected LLMs in the creation of annotated data that closely resembles the data produced by humans for each dataset domain across two languages;
- Compare the model's performance when trained on the original dataset versus when trained solely on synthetic datasets. This comparison should assess the synthetic data's effectiveness and whether it is representative in terms of capturing the diversity and the linguistic variation (tone) of actual commentators;
- Train a stance detection model using both synthetic datasets and existing real-world data to see whether incorporating the synthetic data improves the model's performance, particularly in domains where labelled data is limited. This can be seen as the primary objective of this study;
- Determine the challenges faced by LLMs with respect to generating new data in order to simulate real-world data.

1.5 Road Map

In addition to **Chapter 1**, which introduces the overall outline and purpose of the dissertation, this study consists of five further chapters. **Chapter 2** provides a background to the main concepts relevant to the dissertation and the related works that demonstrate LLMs' contributions in previous studies. **Chapter 3** presents the overall approach, utilising both ChatGPT and Falcon to produce synthetic datasets before testing the performance of these data within various stance detection models. **Chapter 4** details the implementation process, including dataset description, prompt strategies, pre-processing, and model selection. **Chapter 5** then discusses the results of various experiments on the generated datasets, to address the study's main objective. Finally, **Chapter 6** offers some conclusions, discusses the limitations of the study, and offers recommendations for future work.

Chapter 2: Literature Review

2.1 Overview

This chapter is divided into two main sections: **Background** and **Related work**. In the first section, we will introduce the underlining concepts, including stance detection, transformer, LLMs, and synthetic dataset, to enhance the understanding of the following chapters. The second section presents previous works showcasing the approaches employed for stance detection, the LLMs’ capabilities in annotating and augmenting the datasets, and the common prompts strategies used to guide these models.

2.2 Background

2.2.1 Stance detection

Stance detection is a relatively new field of research in NLP that relies on the concept of extracting information automatically from a text (Alturayef et al. 2023). Stance detection aims to determine the stance of the author of a text, commonly discussed in terms of whether the author is in favour of the subject under discussion, against it, or neutral with respect to it. Such topics may be discussed across several different platforms, such as debate websites, news commentaries, and social media (Alturayef et al. 2022). For example, given the text: *“We live in a sad world when wanting equality makes you a troll”*, and the target: “Feminism Movement”, the stance toward the target should be set to “Favour”. While this may be clear to a human reader, finding an effective means of implementing such classification automatically could play a vital role in supporting decision-making for businesses, governments, and authorities by helping provide valuable insights into public opinions more readily (Alturayef et al. 2023).

Research efforts in this area have involved collecting data for diverse targets (e.g., “Hillary Clinton” ,“Feminism”, “Climate Change”, etc.), as shown by (Mohammad et al. 2016a). Some such endeavours have also extended beyond the boundaries of the English language, incorporating languages such as Arabic (Alturayef et al., 2022). Overall, X, formerly known as Twitter, has been the most commonly investigated platform within such work due to its

relatively uncomplex API¹ as compared to other platforms and its ability to provide large volumes of real-time data on a wide variety of topics.

The majority of stance detection datasets can be categorised according to the target into in-target, cross-target, and target-independent (Alturayef et al. 2022). In in-target stance detection, as shown by Mohammad et al. (2016a), the annotation process focuses on a specific entity or topic (e.g., “Hillary Clinton”) that is explicitly mentioned in the text. This enables stance detection models to be trained and evaluated for those specific targets. In cross-target stance detection (Augenstein et al. 2016), such detection extends to a broader range of new targets, allowing for a more generalised understanding of patterns to identify stances of targets not seen in the training phase. Target-independent stance detection (Bar-Haim et al. 2017), on the other hand, focuses on detecting a stance towards a claim or news without relying on specific target annotations. This approach allows for the development of stance detection models applicable to a wide range of contexts without the need for target-specific annotations. In this manner, it overcomes the constraints of in-target and cross-target detection caused by the focus on a target as an explicitly mentioned entity. This dissertation, however, focuses on in-target stance detection, where the target will be specified in the training and evaluation processes.

The main challenges observed in the revised literature for stance detection are in the annotation process. Annotating a large dataset is often time-consuming and costly due to the manual effort required to read and analyse each instance before the assignment of the most appropriate labels. Furthermore, as observed by Alturayef et al. (2022), in-target stance detection annotations must be specific to the target of interest, whether that is “women’s empowerment”, “Covid-19 vaccines”, or “digital transformation”. Hence, annotating data for multiple targets requires additional effort, as well as the application of expertise to ensure accurate annotations for each target. It is also often difficult to achieve high inter-annotator agreement due to subjectivity, as annotators may have different stance judgments with respect to a single text. This could happen because of some ambiguities or the need for a deeper understanding of the subject matter, which may not always be available among annotators.

As a result, these challenges tend to lead to data scarcity in the form of limited datasets or imbalances between classes, potentially affecting the performance of stance detection models.

¹ Application Programming Interface is a set of rules that enables communication and data exchange between software applications.

2.2.2 Transformers

The Transformer model architecture was introduced by Vaswani et al. (2017). It relies on an attention mechanism to resolve long-range dependencies limitations (i.e., pairs of words or phrases that are far apart in meaning, but are still related) for sequence-to-sequence tasks. This attention mechanism allows the Transformer to capture the context of each word and its relationship to other words in the sequence by assigning different scores for the tokens, indicating the importance of each word in the sequence. A Transformer mainly consists of an encoder, which processes the text into meaningful representations, and a decoder, which creates the desired output using the input's representation. By 2018, Transformers became the default architecture for a wide range of language models, including Generative Pre-trained Transformers GPT (Radford et al. 2018) and Bidirectional Encoder Representations from Transformers BERT (Devlin et al. 2018). BERT utilises a bidirectional encoder to allow the development of a better understanding of the context. It was pre-trained using a massive unlabelled dataset drawn from Wikipedia and a range of books in Google to undertake both Masked Language Modelling (MLM) and Next Sentence Prediction tasks.

In MLM tasks, some random input tokens are masked, with the model then being asked to predict the real tokens based on the input context as a way to train the model to understand the relationships between words. The Next Sentence Prediction task can then be utilised to train the model to understand the relationships between sentences; it does this by providing two sentences and asking the model to predict whether the second sentence complies with the ideas within the first sentence. Due to BERT's success in various NLP tasks, such as text classification, several new versions have been created, such as RoBERTa (Liu et al. 2019) and AraBERT-twitter (Antoun et al. 2020), and these offer various enhancements. Such BERT-base models have thus been fine-tuned for the stance detection task used in the current study.

2.2.3 Large language models

LLMs such as OPT (Zhang et al. 2022), GPT-3 (Brown et al. 2020), and ChatGPT have demonstrated impressive text-generation capabilities, leading to a revolution in NLP. These models typically consist of a minimum of one billion parameters and are commonly trained on massive corpora of text drawn from internet resources, such as Wikipedia articles, books, and web crawls (Pahune and Chandrasekharan 2023). For example, GPT-3 was trained on 500B tokens with 175B parameters, while OPT was trained on 180B tokens with a total of

175B parameters. This extensive training, supported by a sizable Transformer model, allows such LLMs to effectively understand and generate human-like text (ibid).

According to Min et al. (2021), LLMs can be fine-tuned to adapt to specific tasks, such as determining semantic similarity, performing sentiment analysis, or text classification. Alexandr et al. (2021), for instance, fine-tuned and adjusted the hyperparameters of a GPT-3 model for text summarisation for Russian texts. Another way to tackle specific NLP tasks with LLMs is the use of a prompt-based approach, where instructions are given to the model to generate the desired output without updating the model's parameters (Min et al. 2021). This was seen in Zhang et al. (2023), who employed several different LLMs such as Flan-T5, text-003, and ChatGPT to determine the expressed sentiment (i.e., positive, negative, or neutral) in a given text. Such experiments highlighted the ability of LLMs to provide comparable performance to a fine-tuned model with respect to sentiment analysis tasks.

LLMs have also had their capabilities significantly expanded with respect to NLP tasks outside of English language settings in recent years. For example, Winata et al. (2021) highlighted the ability of GPT models such as GPT-2 (Radford et al. 2019) to make predictions using several different languages, including German, French, and Spanish, based on few-shot learning without the need to fine-tune the model.

Although LLMs show impressive capabilities with respect to producing coherent and contextually relevant text, they commonly lack a deep understanding of the fundamental concepts being discussed (Farrokhnia et al. 2023). Hence, they may provide inaccurate or nonsensical responses, particularly in complex or ambiguous scenarios.

The next two sections look more specifically at the two LLMs used in the current study.

2.2.3.1 ChatGPT

A Generative Pre-trained Transformer (GPT) is a series of LLMs created by OpenAI. These models are based on Transformers architecture and pre-trained on massive unlabelled datasets, allowing the model to create human-like text. GPT models have shown an exceptional ability to attain a more advanced understanding of language with a score of 45.4 in the Corpus of Linguistic Acceptability (CoLA) task², easily surpassing the other state-of-the-art models at 35.0 (Min et al. 2021).

² The CoLA task (Warstadt et al 2019) is designed to evaluate language models' capabilities with respect to understanding syntax and grammar, as well as their ability to produce coherent texts.

ChatGPT is part of the GPT family of LLMs, which is based on the GPT-3.5 series. It was trained using Reinforcement Learning from Human Feedback (RLHF) in which it received feedback and guidance from human trainers and then worked to “align” its behaviour to that user’s queries, generating responses focused on attaining the desired outcomes (OpenAI 2023a). Multiple layers of feed-forward and self-attention neural networks are thus used to construct ChatGPT (Hassani and Silva 2023). These self-attention networks allow the model to assign different weights to different words to capture dependencies between words, while the feed-forward processes take information from the self-attention layers to generate meaningful representations of input (Vaswani et al. 2017). Based on this process, ChatGPT has a powerful capability to generate coherent and relevant text responses.

With the ability to produce high-quality human-like text, ChatGPT has recently contributed to various applications. One such application is question and answering in different fields such as education and medical fields, where ChatGPT has shown an impressive range of capabilities in addressing mathematical problems and answering patients’ questions (Liu et al. 2023). Additionally, ChatGPT has been used to generate text in various forms, including making inferences based on effectively grasping the intended message within a text, crafting commentary articles, as well as writing code (ibid). Another task is text classification, where ChatGPT makes immense efforts in tackling challenges such as sentiment analysis using its ability to predict the text class accurately with zero-shot learning (i.e., where the model can perform a specific task without explicit training) or few-shot (i.e., train the model with a small number of examples) learning manner (Zhang et al. 2023).

According to Liu et al. (2023), ChatGPT can provide results that are comparable to, or even better than, the results from fine-tuned models such as BERT and GPT-3.5 in terms of zero-shot performance. However, it is important to note that it has limitations with respect to complex reasoning tasks such as non-textual semantics (mathematical reasoning) and neutral similarities in texts.

2.2.3.2 Falcon:

Falcon is an open-source LLM released by the Technology Innovation Institute (TII) in 2023 (Almazrouei et al. 2023). It was launched in two versions: Falcon 40B, which was trained on 1 trillion tokens with 40 billion parameters; and Falcon 7B, which was trained on 1,500 trillion tokens with 7 billion parameters. The data used to train Falcon was the RefinedWeb

English dataset, a high-quality dataset created by TII based on web crawls of public content (Penedo et al. 2023). According to Ferrag et al. (2023), after its release, Falcon moved to occupy first position on the Open LLM Leaderboard³ by June 15, 2023, surpassing other comparable open-source models on the Hugging Face Hub, such as RedPajama⁴ and MPT-7B⁵. In this study, we used Falcon 7B due to memory limitations, as the other version consumes well over 16 GB of memory.

The Falcon LLM (Almazrouei et al. 2023) is an autoregressive decoder-only model that relatively accurately predicts the next token in a sequence while using fewer computationally resources as compared to GPT-3. As stated on the Falcon official website, the Falcon model can be used for a range of NLP tasks, including text generation, question-answering, summarisation, and code-generation.

2.1.4 Synthetic Dataset

A synthetic dataset is any computer-generated dataset that has been annotated artificially through algorithms or simulations to mimic the characteristics of real-world data (Jacobsen 2023). In other words, generating such a dataset involves having similar statistical properties, distributional patterns, and relationships that appear in real data.

Generating synthetic data using LLMs such as ChatGPT helps address various problems in many NLP tasks. One common issue is data scarcity, which can impact the training and evaluation of NLP models. This occurs, for example, in the stance detection task, where most of the datasets are domain-specific, covering contentious topics such as climate change (Upadhyaya et al. 2023), and US elections (Grimminger and Klinger 2021). Consequently, this domain specificity results in data scarcity issues due to the costly and time-consuming nature of manual annotations for new uses. Generating new synthetic data thus allows models to address various topics such as healthcare and finance; this is useful when there is limited access to real-world data or when the sensitive information cannot be shared due to privacy concerns (Nikolenko 2021). Additionally, authentic datasets may be too limited or biased for particular applications, which may lead the models to overfit and fail to generalise well (ibid).

³ A platform for monitoring, rating, and evaluating different language models and chatbots as researchers create and release them.

⁴ [togethercomputer/RedPajama-INCITE-7B-Base](https://togethercomputer.com/RedPajama-INCITE-7B-Base) · Hugging Face

⁵ [mosaicml/mpt-7b](https://mosaicml.com/mpt-7b) · Hugging Face

In such cases, augmenting existing datasets with synthetic data can help increase the accessible dataset size, mitigating several issues.

Synthetic data usage has recently increased in various fields of machine learning, including computer vision and NLP applications. However, various challenges have emerged in terms of using synthetic data to improve real-world applications (Hassani and Silva 2023). The main concern is the need to guarantee the quality of the synthetic data to ensure that it is representative of the original data and captures the proper structure, a process which may be difficult due to textual ambiguity. Moreover, if the generation process is not carefully designed or the original training data itself is biased, synthetic datasets can inadvertently introduce further biases that may lead the model to offer unsatisfactory performance.

In summary, generating synthetic data offers a cost-effective solution to data scarcity in various tasks and could improve model performance by creating more representative and diverse datasets.

2.3 Related Works

2.3.1 Automatic stance detection

The dataset is one of the core features of the different approaches used to tackle stance detection tasks. Substantial effort has been devoted to creating annotated datasets. For example, Mohammad et al. (2016a) formulated stance annotation guidelines to construct well-labelled datasets; other studies have followed these guidelines to develop new datasets, as seen in Glandt et al. (2021) and Alturayef et al. (2022). Most stance detection datasets are publicly available, and they span various text types (e.g., tweets, news headlines, or posts in online forums) and multiple different languages, including Chinese, Arabic, English, and Turkish. A comprehensive overview of published datasets can be found in the work of Küçük and Can (2020) and Alturayef et al. (2023).

Various methods have been utilised in studies for in-target stance detection tasks. One common method is supervised-based learning, where a model is trained on an annotated dataset, with its performance then evaluated on unseen examples (Alturayef et al. 2023). This learning system can be classified into three approaches: traditional machine learning models, deep learning algorithms, and ensemble approaches. Traditional machine learning models (e.g., Logistic Regression) depend on extracting features such as n-gram and

sentiment dictionaries before training the classifier. Mohammad et al. (2016b) used a Support Vector Machine (SVM) with characters and word n-grams as features. They trained their model using their SemEval-2016 dataset achieving an average F1 score of 68.98 for the “Favor” and “Against” classes.

Another technique, deep learning algorithms, uses a complex neural network such as a recurrent neural network (RNN), convolutional neural network (CNN), or transformer (e.g., BERT) to capture complex patterns from text. For example, Long Short-Term Memory (LSTM), a type of RNN, has been commonly used in more than ten studies, and usually shows favourable results for this task as it can capture long-term dependencies in the texts (Küçük and Can 2020). Transformers, also frequently employed in recent studies, demonstrate remarkable performance. Cotfas et al. (2021), for instance, fine-tuned the BERT model by setting the appropriate hyperparameters and then trained it on a pre-processed labelled dataset created for the Covid-19-vaccine target. The test results showed 78.94 accuracy, suggesting that it outperforms models such as SVM. In ensemble approaches, multiple classifiers are combined using various strategies, such as majority voting, to produce a final stance prediction for a text, as presented in Siddiqua et al. (2018).

According to Alturayef et al. (2023), deep learning algorithms demonstrate better performance than traditional machine learning models in various studies. This dissertation thus utilised a deep learning approach, specifically using BERT-base models applied to existing datasets from different sources, with the results presented using a range of different metrics.

2.3.2 Data augmentation using LLMs

Data augmentation, or data enhancement, is one of the most common techniques used to generate synthetic data. This technique helps address the challenges seen in various NLP tasks where limited annotated data is available or when the number of examples across different classes is imbalanced. As Feng et al. (2021) state, this process aims to increase the number of data examples and the data variety of an existing dataset to make it more representative of the actual data without the need to collect more.

Several methods have been proposed for data augmentation. Wang et al. (2022) presented three approaches derived from the lexical substitution method, a technique frequently used to replace words in a text while maintaining overall meaning. One common approach is to

substitute a random word within a sentence with its synonymous counterpart using a database such as WordNet (Miller 1995); this method is based on a synonym dictionary. The second approach is based on the use of a word vector that allows replacing a word with its adjacent word in the embedding space using a pre-trained word embedding such as GloVe (Pennington et al. 2014). Finally, the third approach is based on a language model for substitution by predicting the masked word based on the text’s contextual meaning; this approach’s new text exhibits a higher level of grammatical coherence than the previous methods.

Back-translation is another widely used method for data augmentation (Liu et al. 2020). Many NLP applications, such as summarisation and question-answering tasks, utilise this approach (Feng et al. 2021). In this method, the text is first translated into another language (e.g., Chinese from English), and then translated back into the original language. However, it is worth noting that the accuracy and precision of the translation process determine the quality of the data generated through such back-translation (Liu et al. 2020).

The recent remarkable success of LLMs such as ChatGPT with respect to text generation has motivated researchers to leverage these models for data augmentation to overcome the accuracy and diversity limitations of previous methods. Dai et al. (2023) proposed a new method, AugGPT which utilises ChatGPT to produce additional training data for few-shot text classification to handle the limited amount of labelled data. AugGPT’s performance was tested on datasets from both a general domain and a specific domain and, compared with the other data augmentation methods, AugGPT achieved remarkable double-digit enhancements in terms of accuracy in sentence classification tasks. Accordingly, ChatGPT may be expected to enhance the diversity of the generated data due to pre-training on an extensive corpora, which allows it to have a wider range of semantic expressions to the original data (ibid). Moreover, by employing RLHF, ChatGPT can evaluate the advantages and disadvantages of various expressions to produce high-quality generated data.

Another study (Møller et al. 2023) conducted experiments using ChatGPT and GPT-4 to address the problem of low-resource tasks such as hate speech, sentiment analysis, and social dimensions⁶. A small base dataset of human-annotated data was used to generate a substantial volume of labelled data, based on prompting LLMs to resemble the base set’s examples closely. The results demonstrated that LLMs can create high-quality labelled data that can

⁶ Social dimensions encompass a range of aspects and attributes within social conversations, including themes such as identity expression, support, fun, conflict, etc.

effectively train models in low-resource scenarios, with even a simple prompt. Also, the generated data offers comparable performance to real data for certain tasks, as observed in the social dimensions task, which demands more detailed prompts (ibid).

To sum up, ChatGPT's capabilities demonstrate comparable performance to previous approaches to data augmentation and may thus claim to offer pseudo data that simulates real-world data. Employing a well-designed prompt strategy may also produce more informative and representative data. In this dissertation, attempts are also made to leverage ChatGPT for generating synthetic data; however, in this case it was used to create new data rather than rewrite real data.

2.3.3 Annotation process using LLMs

Several studies have showed the ability of LLMs (specifically from the GPT-3.5 series, such as ChatGPT) to annotate data, thereby addressing the costs associated with manual data labelling. For example, Gilardi et al. (2023) demonstrated that ChatGPT could generate better annotations than Amazon Mechanical Turk (MTurk)⁷ crowd-workers in terms of zero-shot classification tasks. The experiments were conducted on a dataset of tweets labelled for five tasks: relevance, topic, stance detection, general frame, and policy frame detection, each with a different number of classes. Their results showed that ChatGPT outperformed MTurk in terms of accuracy in four tasks, exhibiting good agreement with the original labels, while in certain tasks (including stance detection), the resulting accuracy was up to 3.4 times greater than that of MTurk. As noted by Zhang et al. (2022), ChatGPT utilising a zero-shot prompt strategy in stance detection, also achieved results comparable to other state-of-the-art models trained on 80% of the training data. Accordingly, ChatGPT appears to offer a satisfactory and cost-effective solution for data annotation even without training. Furthermore, unlike other classifiers, ChatGPT can provide reasons for selecting a class (ibid).

The annotation process often involves defining the task, the meaning of each label, and providing examples as a reference to annotators. He et al. (2023) thus proposed AnnoLLM as a general-purpose data annotator system that simulates the manual annotation process as a way to guide GPT-3.5 with its data labelling. To improve the performance of LLM annotators to surpass or match manual annotators, they suggested an “explain-then-annotate” approach.

⁷ Murk is a platform for online crowdsourcing run by Amazon. It allows academics to outsource simple human intelligence tasks such manual data labelling.

This approach utilizes few-shot learning and chain of thought (CoT) (Wei et al. 2022) to provide examples for each label alongside a rational explanation of the reasons why the text is labelled as it is. Compared to the GPT-3.5 annotator in zero and few-shot settings, the few-shot CoT setting offers accuracy improvements of 6 to 8 figures.

The advantages of ChatGPT in terms of annotating data suggest that ChatGPT offers a compelling alternative solution for creating labelled data in various NLP tasks. It can even outperform manual annotators, and its ability to provide reasons for its labelled classes further increases its effectiveness. It was thus decided to test ChatGPT, and other LLMs in creating an annotated dataset for the stance detection task in this study.

2.3.4 Prompt-based learning:

Models based on the GPT architecture, such as ChatGPT, are designed to respond to prompting rather than fine-tuning. Prompting refers to providing instructions in text form, often short sentences, to a pre-trained model to encourage that model to generate a response based on knowledge acquired during training (Yuan et al. 2021). One of the advantages of using prompts is the ability to reduce computational demands compared to fine-tuning methods, by eliminating the need to update the parameters of the pre-trained language model (Min et al. 2021). Additionally, prompts can more effectively guide the model to utilise the knowledge acquired during pre-training to perform a specific task. For instance, when a user interacts with ChatGPT using a prompt such as “Write an essay about climate change”, the model employs its existing knowledge to create output that aligns with the theme given in the prompt without fine-tuning its parameters.

Many studies show various types of prompting help to guide a model to the desired output. This work thus presents two common approaches: template-based prompts and learning from instructions (Min et al. 2021). Template-based prompts involve placing templates or structures containing specific slots to encourage a model to generate responses by filling the spaces with appropriate information based on the context (ibid). For example, Wang et al. (2022) used cloze-style prompts to design a template for a stance detection task that used just the target and stance slots in the form “The stance towards {t} is [MASK]”, where {t} is used to predict the target and MASK is used to predict the relevant label (i.e., favour, against, or none).

Unlike template-based prompts, learning from instructions can provide greater detail and describe the task more explicitly. Zhang et al. (2023) utilized this approach to evaluate various LLMs' performance in sentiment analysis tasks in both zero-shot and few-shot classification settings. The prompts for zero-shot learning were simple and included the essential elements, such as: the task name; the task definition to provide additional guidelines and the labels options for the model; the output to establish the expected format for the response; and a demonstration offering an additional element in few-shot learning to provide examples with labels for each class. An example is as shown in Figure 1. Their findings demonstrated that, in zero-shot scenarios, LLMs offered satisfactory performance on simple tasks, while in few-shot scenarios, LLMs steadily outperformed the finetuned model.

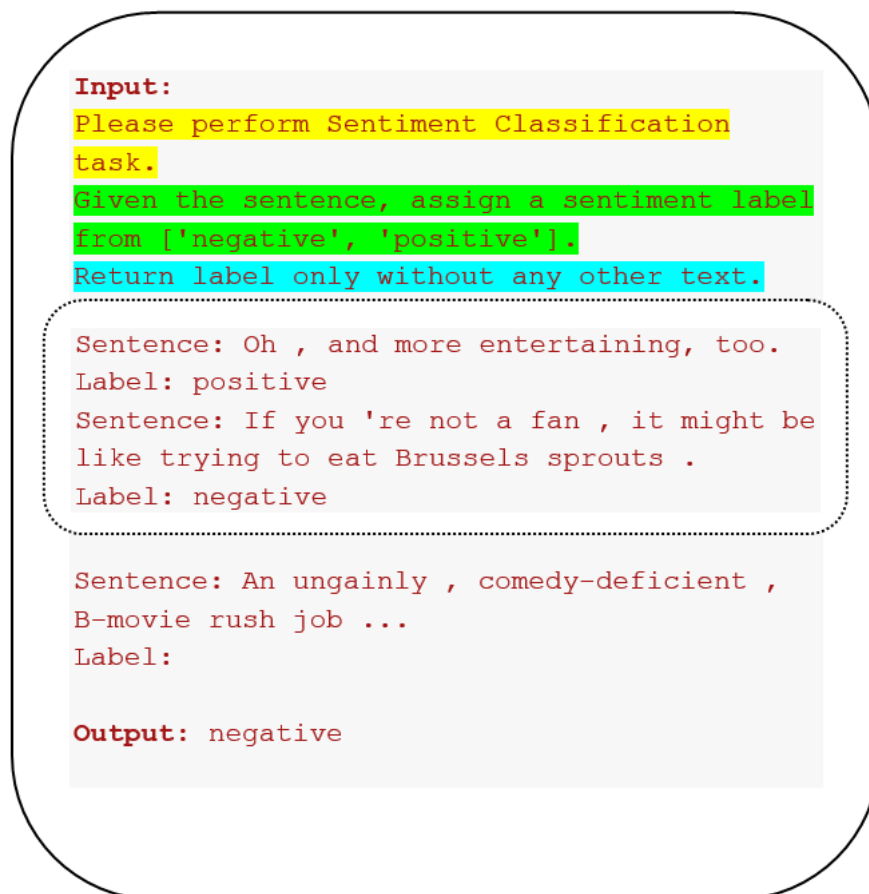


Figure 1. A prompt by Zhang et al. (2023). It includes the task name (highlighted in yellow), task definition (highlighted in green), and the output (highlighted in blue). The demonstration examples (surrounded by dots) operate in few-shot settings; these would be removed in zero-shot settings.

To assess the significance of prompt design, Zhang et al. (2023) used GPT-4 to generate additional prompts for sentiment analysis tasks and compare their performance with manually created prompts. The results suggested that as prompt choice may vary depending on the

nature of the task and the desired output, which means the tasks with less structured output cause the prompt choice to have less effect on the model's performance.

One of the main challenges in generating data with simple prompts is faithfulness, that is when the patterns and characteristics of the synthetic data fail to match the original data. Veselovsky et al. (2023) proposed different strategies to address this problem. Of particular interest for this dissertation is the ones called grounding prompt. This technique uses an example drawn from real data in the prompt, asking the model to generate similar examples or to transfer the style. This strategy shows best performance because it encourages the model to absorb the data patterns and language styles present in the data and not be biased towards the types of data it was initially trained on.

It seems, therefore, that template-based is less flexible, as these rely on predefined structures. On the other hand, instruction prompts offer greater flexibility but may lead to less precise results or a tendency to overgeneralise. Including real-world data in prompts may thus be seen as necessary to effectively direct the model to produce the intended results. In this dissertation, a combination of these approaches is employed to motivate the LLMs when generating effective new data. Instruction-prompt choices are thus used to create the context, while template-based approaches help construct that context with the appropriate stance.

2.4 Summary

The literature review presented in this Chapter discussed the frequent approaches employed for tackling stance detection tasks, focusing on supportive-based learning methods. It also presented LLMs' contributions across different NLP tasks, showing their impressive capabilities. These abilities are observed in the augmentation process, which increases the datasets while preserving semantic similarity. LLMs also exhibit effectiveness in data annotation, providing comparable performance to human crowd-workers. These contributions make LLMs a potential resource for creating new annotated datasets. Additionally, the review explored the significance of diverse prompting strategies for LLMs.

Chapter 3: Methodology

This chapter briefly presents the approach used to investigate the LLMs’ ability for generating new data in ways that may improve model performance for stance detection tasks across different languages and topics.

3.1 Datasets and models preparation

Initially, existing manually annotated datasets for stance detection tasks were investigated. Several datasets in different languages centred around the same topic were identified in order to evaluate the performance of ChatGPT and Falcon in various linguistic contexts. In English, the Feminist Movement topic (Mohammad et al. 2016a) was selected, while in Arabic language, Women’s Empowerment (Alturayeif et al. 2022) was selected. To extend the study to another topic in the English language, the Covid -19 Vaccine dataset (Cotfas et al. 2021) was chosen. All selected datasets were collected from the X platform (formerly known as Twitter). A more detailed description of the datasets is provided in Chapter 4.

The LLMs were then used to generate synthetic tweets to simulate the original datasets. For the ChatGPT model, specifically gpt-3.5-turbo, we employed various prompt strategies to create tweets with different perspectives in the target domain over a range of possible tones. Three approaches to creating tweets were investigated:

- **General and simple prompts for all stances**

Initially, a prompt for the model was created to generate labelled tweets related to the target (e.g., feminism) in the form “Tweet. [stance]”. However, these generated tweets were not accurately annotated; this approach was thus abandoned.

- **Specific prompts for each stance**

In this approach, specific prompts were designed for each stance label (favour, against, neutral) with respect to the target topic.

- **Template prompt for each stance**

Following on from the concept of designing a prompt for each stance, prompts were constructed with hint sentences in the form “Template. [topic].#hashtag”, as a way to cover a wider range of subtopics discussed in the target domain, as observed in the real data.

The final synthetic tweet dataset created using the ChatGPT model was produced by employing both the **Specific prompt for each stance** and the **Template prompt for each stance** approaches. For the Falcon LLM, the Falcon-7B- instruct model⁸, which is a fine-tuned version of Falcon 7B, was utilised. This model was customised to provide suitable responses to prompts in an instructive manner. For this model, simple prompts consisting of a task description and the output format were designed, based on the same concept as the **Specific prompt for each stance** approach.

At the end of this phase, five synthetic datasets were created using the two different LLMs. These datasets were then used in the experiments to observe the performance of stance detection models with each synthetic datasets individually.

In the process of choosing stance detection models, reference was made to the research papers that created the original datasets. For the Feminism and Covid-19 topics, the RoBERTa⁹ and Distilbert¹⁰ models, in different sizes in English were used, as the Feminism model was not publicly available, while AraBERT-twitter¹¹ was used for Arabic. All these models are available on the Hugging Face Hub.

3.2 Evaluation

After generating the dataset and selecting stance detection models, multiple experiments were conducted to train the models for each topic via the employment of real, synthetic, and merged (i.e., those containing real and synthetic data) datasets. A detailed description of these experiments is provided in Chapter 5. The performance of each model was evaluated via artificial and real test sets, and the results were presented in four metrics: Accuracy, Precision, Recall, and F1 score. These metrics are defined using the following values: True Positive (TP), which indicates the number of tweets correctly predicted by the model as belonging to a specific stance class (when they explicitly belong to that class); True Negative (TN), which denotes the number of tweets correctly identified as not belonging to a specific stance class (when they do not explicitly belong to that class); False Positive (FP) indicates the number of tweets that the model incorrectly classified as belonging to a specific stance class (when they do not explicitly belong to that class); False Negative (FN) refers to the

⁸ [tiiuae/falcon-7b-instruct · Hugging Face](#)

⁹ [cardiffnlp/twitter-roberta-base-sentiment-latest · Hugging Face](#)

¹⁰ [distilbert-base-uncased · Hugging Face](#)

¹¹ [aubmindlab/bert-base-arabertv02-twitter · Hugging Face](#)

number of tweets incorrectly predicted as not belonging to a specific stance class (when they explicitly belong to that class). Cotfas et al. (2021) defined each metric as follows:

- **Accuracy**

Accuracy is considered to be suitable for a balanced dataset by providing an overall measure of the model performance. It calculates the ratio of correct predictions to the total predictions made for the dataset:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

- **Precision**

Precision represents the ratio of true positive predictions to the total number of tweets that classified as belonging to a certain class. Mathematically, it can be defined as:

$$\text{Precision} = \frac{TP}{TP+FP}$$

- **Recall**

This metric establishes the proportion of true positive observations out of all actual tweets belonging to a certain stance class in the dataset. Mathematically, it can be defined as:

$$\text{Recall} = \frac{TP}{TP+FN}$$

- **F1 score**

The F1 score is suitable for an imbalanced dataset such as the Feminism dataset. It is calculated as the harmonic mean of precision and recall:

$$\text{F1 score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figure 2 illustrates the overall approach adopted by each of the LLMs (ChatGPT or Falcon) to generate data and train the selected models for stance detection tasks. Additionally, it provides a concise overview of the experiments conducted to understand the performance of each LLM.

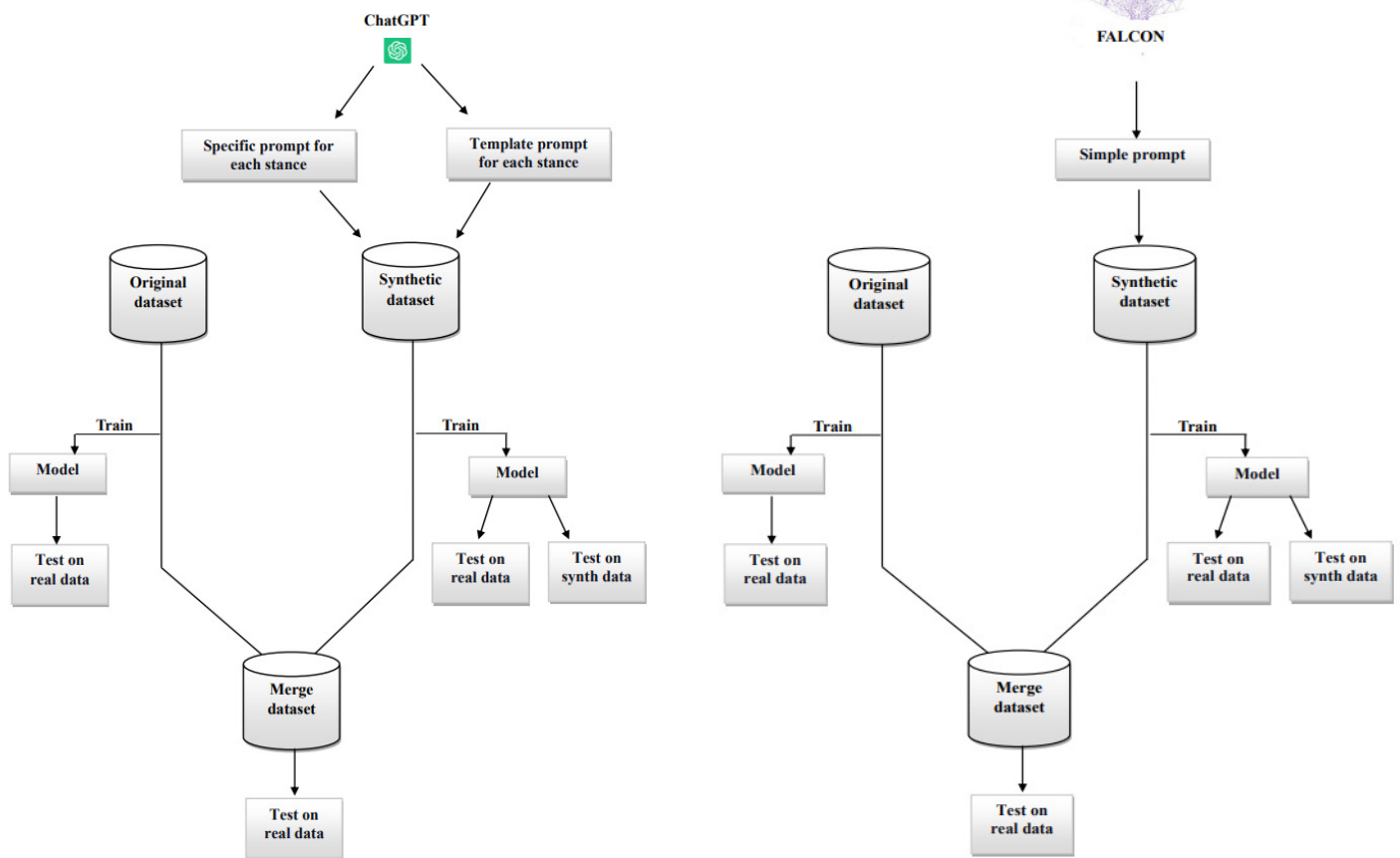


Figure 2. Overall approach with each of the LLMs

Chapter 4: Implementation

4.1 Overview

This chapter outlines the phases undertaken to address this study’s research objectives in detail, along with the justifications for this approach. First, the original datasets employed in this project are described (Section 4.2). Next, the prompt strategies used for each of ChatGPT and Falcon to encourage them to create synthetic data that simulates real datasets as closely as possible are discussed (Section 4.3). The pre-processing phase used for both original and synthetic datasets is then explained (Section 4.4), followed by a discussion of the models selected for the stance detection task (Section 4.5) which were primarily used to evaluate the experiments conducted with both original and generated datasets.

4.2 Dataset descriptions

One of the objectives of this study was to create new synthetic datasets using LLMs to increase the training data available for stance detection models. The aim was thus to assess the capability of LLMs to generate data in different languages on various topics. The Feminist Movement and Women’s Empowerment datasets were selected as representing the same domain topic in different languages. The Covid-19 Vaccine topic was additionally selected to assess model abilities in the context of another theme. All selected datasets were publicly available and free for utilisation in a research context.

- **English language data**

The Feminist Movement dataset is a subset of the SemEval-2016 Stance Dataset collected and annotated by Mohammad et al. (2016a) in their work on stance detection tasks. The Stance dataset consists of 4,163 labelled tweets across five diverse targets: Atheism, Climate Change, the Feminist Movement, Hillary Clinton, and Legalisation of Abortion. The Feminism subset contains 949 tweets, which were originally divided into a training set of 664 (70%) tweets and a test set of 285 (30%) tweets, with the latter used to assess a model’s ability to generalise to unseen data. This data was organised into two columns: text of the tweet and stance of the tweet. The stance might be “Favor”, in support of feminism; “Against”, criticising feminism, or “None”, which represents a neutral stance or indicates a lack of clarity in the tweeter’s stance towards feminism. The tweet distribution among the

stances was assessed as follows: 210 tweets with the Favor label, 328 tweets Against, and 126 tweets with the None stance; demonstrating an imbalance between classes.

Another dataset in the English language, the Covid-19 Vaccine data, was created by Cotfas et al. (2021). Due to a policy implemented by X (formerly known as Twitter), the annotated dataset contained tweet IDs rather than tweet texts, requiring the use of the Twitter Academic API to restore each tweet's text. The annotated dataset consisted of 3,249 tweet IDs, with an equal number of tweets for each stance. In total, only 2,377 tweets were retrieved, this affected the distributions among the classes. Therefore, the decision was taken to reduce the size of the dataset to 1,887, based on the smallest class, to restore balance to the dataset as seen originally. The final dataset was then split into a training set of 1,604 tweets and a test set of 283, using the original ratio of 80% to 20%, aligning more effectively with the test set size in the Feminism dataset.

- **Arabic language data**

MAWQIF (Alturayef et al. 2022) is an Arabic dataset for stance detection that focuses on three trending topics in the Middle East: Women's empowerment, the Covid-19 vaccine, and digital transformation. Like the Feminism movement dataset in English, the women's empowerment dataset captures opinions on women's rights issues in Arabic culture. The original Women's empowerment dataset consists of 1,400 tweets split into a training set (85%) and a test set (15%). The dataset contains two columns featuring the tweet itself and the stance label, "Favor", "Against", or "None" with respect to women's empowerment. The distribution of tweets by stance was as follows: 760 tweets with the Favor label, 371 tweets for Against, and 59 tweets labelled None.

The datasets were mainly used to conduct experiments, create synthetic data, and train and evaluate the models for each language separately. Tabel 1 summarise the datasets.

Dataset	Language	Target	Train			Test		
			Favor	Against	None	Favor	Against	None
SemEval-2016	English	Feminism Movement	210	328	126	58	183	44
Covid-19 Vaccine	English	Covid-19 Vaccine	535	535	535	94	94	94

MAWQIF	Arabic	Women's Empowerment	760	371	59	134	65	11
---------------	--------	------------------------	-----	-----	----	-----	----	----

Table 1. A summary of the datasets

4.3 Prompt strategies

4.3.1 Prompt for ChatGPT

To simulate real-world tweets, designing prompts that effectively guided ChatGPT in its generation of annotated tweets was important, as the objective was to create synthetic data containing both tweet content and its corresponding stance towards the target. This section explains experiments undertaken to aid the design of appropriate prompts by utilising empirical evidence applying various techniques, including zero-shot and few-shot prompting, as implemented by He et al. (2023) for data annotation.

During experimentation, the ChatGPT model used was the gpt-3.5-turbo1 model drawn from OpenAI. The temperature parameter was set to 1 to allow the creation of more diverse tweets in English. However, for Arabic tweets, it was adjusted to 0.7 to improve coherence, since using a value of 1 did not consistently produce coherence texts.

The data was manually checked to gain insights and to ensure the desired output quality. Prompts with a similar structure were created for both languages and all topics. Additionally, the distribution of generated tweets matched that of the original tweets, allowing for fair and meaningful comparisons. The following sections present the prompt experiments conducted.

4.3.1.1 General and simple prompts for all stance

Drawing inspiration from the prompt used by Michail et al. (2023) to generate small samples of tweets, the initial experiment to generate tweets that incorporated all three stances at once was achieved by using general and simple prompt in the form of introductory sentence for the model, a task description, and the output format, as shown in Figure 3.

Although the produced tweets were satisfactory (Figure 4), such prompts may not produce accurate reflections of the specifics of each stance, which may affect the diversity of perspectives in a given stance. Furthermore, despite a request for equal numbers of tweets for each class, a bias emerged toward tweets that expressed the “Favor” stance, potentially due to the model’s weak performance in terms of tackling negative connotations (Zhong et al. 2023).

Moreover, as Farrokhnia et al. (2023) noted, ChatGPT's lacks deep understanding of the specific context that human beings possess as well as the underlying. This may lead to coherence issues and affect consistency across stances, which could affect the annotation process, as observed in some generated tweets where Favor tweets were incorrectly labelled with the "Against" stance (see tweet 10 of Figure 4).

In a separate experiment, ChatGPT was also used to annotate the original test set. The results showed agreement with the original labels of 73.33%, which almost aligns with the findings of 78.7% determined by Gilardi et al. (2023) in their evaluation of the ChatGPT model's ability to annotate stance. Although this result demonstrates fairly good capability with respect to the annotation process, it is not highly accurate and may introduce errors, making it less suitable for the task of providing synthetic labelled data to enhance model performance. Consequently, a decision was taken to focus on generating tweets for each stance separately in order to obtain more accurately labelled data.

You are a tweet generator.

Task: Create 100 informal tweets related to {TOPIC} with different stance (favour, against, or none) for each tweet. The tweets may include errors slang, emojis, mentions and hashtags. Generate equal numbers of tweets for each stance.

The output: Short Tweet (Maximum of 15 words). (Tweet) : [stance]

Figure 3. General and simple prompt for all stances. The first sentence provides a clear explanation of the task itself for the model, while the Task provides a detailed description of the required tweets. The Output then provides the desired format for the generated tweets.

1. "Strong women, smashing the patriarchy like #FeminismIsPowerful":[favour]
2. "Who needs glass slippers when we have glass ceilings to shatter? #FeminismForAll":[favour]
3. "Not all heroes wear capes, some wear lipstick and fight for equality 🦋 #EmpoweredWomen" : [favour]
4. "Real men support women's rights, it's not just a girl thing! 🗣️ #Feminism4AllGenders":[favour]
5. "Feminism isn't about hating men, it's about loving equality for everyone! #EqualityForAll":[favour]
6. "Sorry, but I can't hear the patriarchy over the sound of my accomplishments! #FeminismWins":[none]
7. "Equal rights don't just magically happen, we have to fight for them! 🦋 #ActivismMatters":[none]
8. "Gender equality is not a trend, it's a necessity for a better world! #ProgressiveMinds":[none]
9. "Feminism means empowering each other, not tearing each other down! 🧡 🧡 #SisterhoodStrong":[none]
10. "Rejecting sexism isn't just a phase, it's a lifelong commitment! 🌙 #FeminismForever":[against]

Figure 4. A sample of generated tweets using General and simple prompts for all Stance approach.

4.3.1.2 Specific prompt for each stance

To address the limitations of the previous approach, the task was simplified for the model by generating tweets for each stance separately, allowing the model to focus solely on generating text without annotation. This strategy made it possible to design more specific prompts covering a wider range of topics and perspectives to gain results that were more representative of real tweets. It was also possible to achieve better results by providing detailed instructions tailored to each specific stance (DAIR.AI. 2023).

A zero-shot prompt technique was employed initially, with a structure similar to that used for the general prompt for each stance. It consists of an introductory clause to prepare the system, a task description, text type, and the output format (see Figure 5 without the Example). In the task description, the type of tweets to be generated based on the stance was included alongside the content that might be included in the tweets, such as hashtags and mentions. For the “Favor” stance, informal tweets in support of the target (e.g., feminism) were requested, while for the “Against” stance, informal tweets that imply opposition toward feminism were desired. For the “None” stance, the model was asked to create casual tweets that might not directly relate to the topic but include the topic’s hashtags as they appear in the original dataset (Feminism and Women’s Empowerment). However, for the Covid-19 topic, the model was asked to create news-like content as presented in most tweets in the real data for the “None” class. This approach helped avoid overlap between the “Favor” and “None” stances, as the model tended to produce supportive texts when asked to create a neutral stance using general hashtags, such as #Feminism, #Covid-19. In terms of text type, as inspired by Michail et al. (2023), the desired length of the tweets (e.g., 20 words maximum) was specified. Additionally, the output format was designed to present the tweet’s text directly, omitting the usual opening response sentence from the ChatGPT model (e.g., “My tweet is:”).

The generated tweets by zero-shot prompt show various hashtags and slangs, yet ignored mentions and focused on narrow topics centred around specific subjects, such as smashing the patriarchy and promoting equality for feminism topics, which decreased diversity, especially across a large number of generated tweets. A few-shot approach was then added, which has the same structure as the zero-shot prompt incorporating three examples for each stance from actual tweets, as illustrated in Figure 5. As a result, the generated tweets covered a wider range of topics within the target domain, including gender stereotypes, combating sexism, etc., as an example of the Feminism target. The few-shot prompt technique,

supplemented with real-world tweets, achieved a better understanding of the diversity of conversations surrounding the topic in the real world.

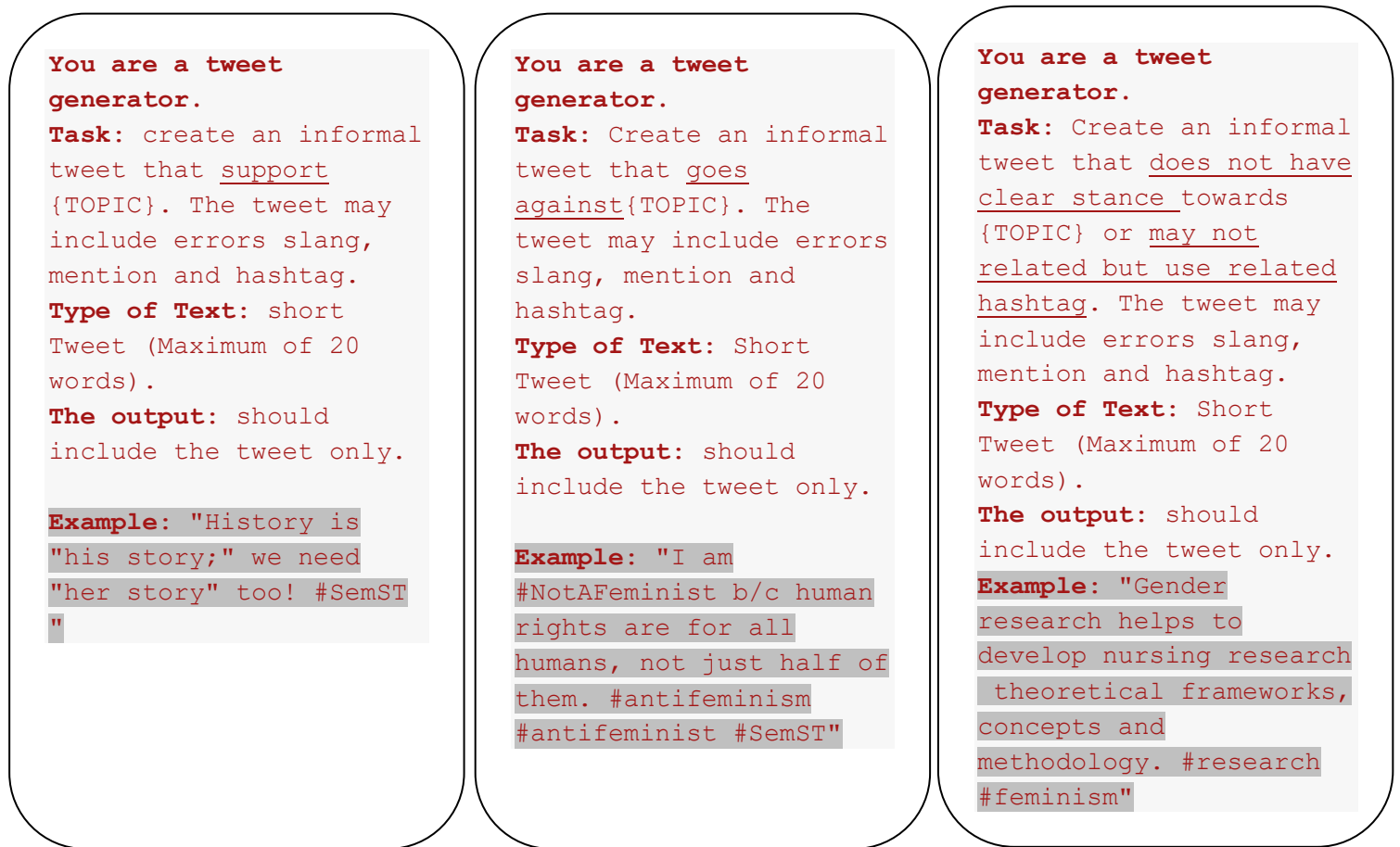


Figure 5. Specific prompt for each stance. “Favor” on the left, “Against” in the middle, and “None” on the right. For tweets in Arabic, the first statement was replaced with: “You are a tweet generator for the Arabic language”, with an Arabic tweet given as an example.

During the experiments, the “n” parameter was set to match the desired number of tweets, which was fewer than 100 in each ChatGPT API call; this encouraged diverse responses for the same prompt to be generated over different runs. Some changes were also made to the examples and slightly modified the task descriptions in each API call, such as requesting the model to create tweets that defend women’s empowerment. These tweaks assisted in increasing diversity and reduced the occurrence of repeated tweets, as each API call operated independently from those in the previous call history.

This approach allowed for the creation of synthetic data with accurately labelled stances, thus avoiding some of the mistakes that tend to arise when combining all stances in a general prompt. However, some semantically repeating tweets still arose across similar API calls.

4.3.1.3 Template prompt for each stance

To push the model to generate a broader range of topics within the selected domain discussed in the original tweets, individual template prompts were designed for each stance (same idea as in section 4.3.1.2). These consisted of the primary sentence for the model, the task description asking the model to generate a tweet for a specific stance using the provided hint, an output format to set the word limit for the tweet and to present only the tweet, examples to demonstrate how to use the hints effectively, and a hint statement in the form “template, [subtopic]. #hashtag”, as shown in Figure 6.

You are a tweet generator.

Task: Create an informal tweet that {stance} {TOPIC}. Inspire your perspective from the hint, and use the exact hashtag correctly.

The output: Short Tweet (Maximum of 20 words) and include the tweet only.

Example: "Hint: you know what's awesome? Feminism!, [reproductive rights for all]. #GenderEquality"

Answer: "Feminism rocks! Let's make sure everyone has access to reproductive healthcare and their rights. #GenderEquality"

Hint:

Hint: "{template} {topic}. {hashtag}"

```
templates = [  
  "you know what's awesome?  
  Feminism!, ",  
  "hey @user, you gotta fight for  
  feminism and ...", ...]
```

```
hashtags= [  
  "#GenderEquality",  
  "#WomenEmpowerment", ...]
```

```
topics = [  
  "[reproductive rights for all]",  
  "[workplace discrimination]", ...]
```

Figure 6. Template base prompt for each stance. In the left, the instruction provided to the model is given with an example. In the right, the Hint statement structure which randomly chooses an element from each list in the form “template, [topic]. #hashtag”

The hint was randomly produced using the elements from three lists: templates, topics, and hashtags. The template list included phrases intended to serve as a basis to create tweets with different perspectives aligned with the selected stance toward the target’s subtopics and in various tones, such as “enthusiastic” and “assertive”, to encourage the model to utilise the subtopic in multiple ways. The chosen phrases were inspired by the original dataset, incorporating styles that were not observed in the generated tweets from the prior approach;

their quantity differs according to the number of stance tweets in real datasets. The topics list included multiple subtopics related to the target mentioned in the original tweets, which ChatGPT may not have captured using the simpler prompt approach (e.g., the Gamergate controversy). Several such topics were extracted using topic modelling (section 4.3.4). The hashtags list was utilised to incorporate the desired hashtags relevant to the target. The lists were coherent each time ChatGPT API was called, creating more coherent tweets. More details for each stance prompt are given in Appendix A.

This approach enabled a more realistic representation of real-world tweets, addressing the issue raised by Alturayef et al. (2022) that some tweets may express a contradictory stance to what is indicated by the hashtag such as “Feminism celebrates women as they are, challenging old-school stereotypes that should be left in the past. Let's keep amplifying voices #AgainstFeminism”. Moreover, this approach allowed ChatGPT to be guided towards generating tweets with more specific topics and tones in a manner that could not be achieved with the previous approach. However, using this type of prompt restricts the model's creativity, producing only the specific style of tweets indicated by the hint; therefore, it is not recommended to employ a template prompt approach for generating all synthetic tweets. This might result in repetitive content and less diverse tweets, which would not align with the aim of this phase of generating broader perspectives. Instead, creating a sample of tweets in this manner and adding these to tweets generated using the process in section (4.3.1.2) is recommended as a way to cover more of the opinions discussed in the original dataset.

To support this opinion, a separate experiment was conducted to evaluate the generated tweets from section (4.3.1.2) and the tweets of this approach. Each synthetic dataset was trained separately using RoBERTa model and tested on the original test set. Synthetic tweets from the previous approach received an F1 score of 56, while the template prompt approach achieved an F1 score of 47. These figures demonstrate that the tweets from section (4.3.1.2) were more representative of real-world tweets. Both synthetic datasets were combined using 30% of template prompt approach data to increase the representative. As a result of this integration, the performance improved slightly to an F1 score of 58.

4.3.1.4 Arabic tweet prompts:

Lai et al. (2023) stated that using English prompts for ChatGPT to generate text in another language can offer better responses than using prompts in the specific language as ChatGPT

is most proficient in understanding English to analyse tasks. Accordingly, we designed prompts in the same structure of Specific prompt for each stance approach and Template prompt for each stance approach using English to generate Arabic tweets for the women's empowerment topic. The only difference was that the provided examples drawn from the original dataset were in Arabic to maintain consistency with the target language. Generated tweets using Specific prompt for each stance show a few minor grammatical mistakes in a small set of tweets. Conversely, more grammatical errors were observed in the Template prompt for each stance approach, as this approach is based on inferring tweets from the hint statement, which may restrict its ability to handle the hint's parts.

Compared to English tweets using Template Prompt, the quality of Arabic tweets was comparatively lower (refer Appendix B for some examples), this highlights ChatGPT's bias toward English language (Lai et al. 2023) . However, overall, the tweets were of good quality and manually checked to remove weak tweets.

4.3.2 Prompts for the Falcon Model:

To create synthetic tweets using the Falcon 7B-instruct model, the previous approach of creating tweets for each class separately was used, this allowed a focus on generating the tweets. However, the model's performance was unsatisfactory for Women's Empowerment in Arabic, possibly due to its training on English and French tokens only (Almazrouei et al. 2023). Therefore, we specifically created tweets about Feminism Movements (in English) and Covid -19 Vaccine topics only.

The tendency was to use simple prompts to guide the model towards generating appropriate data as the model demonstrated inaccurate responses to long and detailed instructions. Hence, a simple prompt consisting of the task and the output format was created, see Figure 7. In the task description the model was asked to create informal tweets with the specific stance toward the specified topic, while the output only required it to generate each tweet in a new line.

To reduce repetitiveness, the prompts description was slightly changed and 20 tweets were generated for each. For example, prompts such as "Create 20 offensive tweets that hate feminism movements" might be used. Additionally, the subtopics extracted from section 4.3.4 were used in the description to push the model towards being more diverse. For instance, this

included prompts such as “Generate 20 informal tweets that criticize workplace *discrimination*”.

```
'''Task: Generate 20 Informal tweets that {support/goes against/ not  
related to} {topic}. The tweets may include errors slang, mention and  
hashtag.  
Separate each tweet by new line. '''
```

Figure 7. Simple prompt for the Falcon 7B-instruct model.

4.3.3 Brief description for the generated data:

At the end of the generation process, three synthetic datasets had been produced by ChatGPT with respect to Feminism Movements, Women’s Empowerment, and the Covid-19 Vaccine. Moreover, two datasets were obtained using Falcon for Feminist Movements and the Covid-19 Vaccine. The synthetic data consisted of 949 generated tweets for the Feminism topic and 1,400 for the topic of Women’s Empowerment in Arabic, matching the size of the original datasets. However, for the Covid-19 Vaccine, only 1,000 tweets were generated opposed to 1,887, with a balanced distribution across all stance labels. These tweets were generated using the Specific prompt for each stance and the Template prompt for each stance approaches in the ChatGPT model. The same volume of data was also produced using the simple prompt in Falcon.

4.3.4 Topic modelling

To explore the possible themes discussed in the original datasets, common topic modelling via Latent Dirichlet Allocation (LDA) (Blei et al. 2003) was employed. LDA is a probabilistic model that assumes that each document is composed of specified topic probabilities, where each topic is a mixture of words with particular probabilities. By analysing the observed words in the document, LDA aims to discover underlying topics and determines topic proportions.

Before training the model, the data were prepared following the approach used by Dahal et al. (2019), which used topic modelling for climate change tweets. To produce meaningful topics, the data was pre-processed by removing numbers, punctuation, usernames, and hyperlinks from the tweets. Additionally, stop words, which are commonly used words that do not add value to the topics, as well as words such as “feminist”, “rt”, “semst”, and “vaccine” that

frequently appeared in the tweets for each topic were removed. The tweets were then converted into bag-of-words representations to simplify the analysis; this ensures that the order of the words is not essential for topic modelling (ibid).

The number of topics was set empirically before training the LDA model by comparing the quality of the generated topics. The model was run with 5 topics, 10 topics, and 15 topics for each stance of tweets. Overall, the 10 topic set performed better than the five topic set, as the higher number of messages provided more insightful information, whereas 15 topics resulted in repeated topics that may not be useful.

The topics were inferred based on their key words. For example, for the Feminism dataset with “Favor” stance we acquired the following keywords for topic 3:

```
(3, '0.013*"sexist" + 0.013*"women" + 0.009*"make" +  
0.009*"misogynist" + 0.009*"views" + 0.009*"expression" +  
0.009*"culture" + 0.009*"got" + 0.009*"feminist" + 0.009*"still"')
```

From the keywords: “sexist”, “misogynist”, and “culture”, the topic [combating Sexism and Misogyny] can be inferred.

The same process applied to the Covid-19 vaccine dataset, as well as the Women’s Empowerment dataset in Arabic:

```
(4, '0.006*"0.004 + "تحقيق"*0.005 + "الله"*0.005 + "السعودية*"URL" +  
0.004*" + "رؤية"*0.003 + "وفي"*0.003 + "حول"*0.004 + "الدولة"  
0.003*"المجالات"*0.003 + "للمرأة"')
```

From the keywords: “رؤية”: vision, “السعودية”: Saudi Arabia, and “للمرأة”: for women, the topic [رؤية السعودية 2030 للمرأة] [Saudi Vision 2030 for Women] can be inferred. Table 2 shows the inferred topics.

	Favor Stance	Against Stance	None Stance
Feminism Topics	[gender equality in various contexts] [combating sexism and misogyny] [putting an end to gender-based violence] [address negative name-calling for women]	[challenges the need for gender equality] [Gamergate controversy] [addressing violence against men] [feminism movements causing harm to society]	[Black Lives Matter] [a wonderful TV show] [friends day!]

Covid-19 Vaccine Topics	[Pfizer-BioNTech vaccine] [the need to get back to jobs] [encouragement for getting vaccinated] [vaccine data improves its benefits] [positive news about the vaccine]	[Not trust in the vaccine] [scepticism about the effectiveness of the vaccine] [vaccine may alter DNA] [scepticism about vaccine development] [vaccine side effects]	[vaccine trials in the UK] [data-related discussions about vaccine] [vaccine availability and health implications] [information about coronavirus in different regions] [vaccine distribution and doses among countries]
Women's empowerment Topics in Arabic	[مساهمات المرأة في التقنية في السعودية] [women's contributions to technology in Saudi Arabia] [رؤية السعودية 2030 للمرأة] [Saudi Vision 2030 for Women] [مساواة المرأة بالرجل] [equality of women with men] [أهمية دعم المرأة في المجتمع] [the importance of supporting women in society]	[نرفض تقليد الغرب في تمكين المرأة] [we reject Western imitation of empowering women] [الوظائف هي حق للرجال] [jobs are a right for men] [واجب المرأة هو خدمة الزوج] [the duty of a woman is to serve her husband.] [الأسر تتفكك بسبب تمكين المرأة] [families break up because of women's empowerment]	[مدعيه النبوه] [claiming prophecy] [هيئة الاعلام والاتصالات] [Media and Communications Authority]

Table 2. Inferred topics for each stance from datasets in English and Arabic

4.4 Data Pre-processing

Before feeding the generated synthetic data into the model, both datasets (i.e., the original and synthetic data) were pre-processed using similar steps to those described in the original papers. The tweets were cleaned by converting all text to lowercase, normalising the mentions to reduce data variability, and removing punctuation and extra white spaces. Additionally, emojis and hyperlinks were removed to ensure the synthetic data was standardised as Falcon contained hyperlinks but no emojis and ChatGPT was the opposite. Subsequently, the cleaned data was encoded using the tokenizer of the selected model for the stance detection task to convert the text into tokens and map each token to a corresponding numerical identifier to be effectively processed by the model.

For Arabic data, the preprocessing steps publicly available from the original data creators were applied (Alturayef et al. 2022). The preprocessing phase began by eliminating non-Arabic letters, diacritics, recurring characters, and tatweel. The tweets were then encoded using the model's tokenizer.

4.5 Model selection

To select the models for the stance detection task, the papers that created the initial real datasets were referenced (Mohammad et al. 2016a; Cotfas et al. 2021; Alturayef et al. 2022) to identify those models that had been fine-tuned to those datasets. Regarding the Feminism movement dataset, a model had to be selected from Hugging Face Hub, as the fine-tuned model mentioned in the original paper was not publicly available. The chosen model was RoBERTa-base, which was trained on a massive corpus of 154 million recent tweets up to the end of December 2022 to update its training (Loureiro et al. 2022). The robustly optimised BERT approach (RoBERTa) (Liu et al. 2019) is built upon the BERT model but incorporates several developments to enhance end-task performance. These improvements include, increasing the training set and batch size, modifying the masking pattern, and augmenting training sequences. In this dissertation, the model was fine tuned for the stance detection task by training it on the Feminism dataset, using similar hyperparameter settings to those described by Loureiro et al. (2022) who evaluated the model's performance on the Stance detection dataset (Mohammad et al. 2016a). However, the number of epochs and the learning rate were adjusted using a cross-validation technique, a process useful with small training sets such as the 664 tweets used here. Furthermore, a smaller BERT-base model from Hugging Face known as the distilbert-base-uncased (Sanh et al. 2019) was selected in order to evaluate the performance of the datasets across a variety of model sizes. This model has fewer parameters (66 million) compared to RoBERTa's 125 million.

For the Covid-19 vaccine dataset, RoBERTa was used in order to focus on the synthetic data sets' quality and to develop a fair evaluation for both English datasets. To fine-tune the model, a training set was employed with hyperparameters set using the validation set. By making these appropriate splits in the dataset, we improved model performance without overfitting the test set.

On the other hand, for the Arabic dataset (Women's Empowerment), the fine-tuning code was made publicly available by Alturayef et al. (2022). The dataset was trained using four different BERT base models attuned to Arabic tokens. On assessing the models' performance with the tuned hyperparameters within the dataset, the AraBERT-twitter model (Antoun et al. 2020) was found to achieve the highest performance, with an F1 score of 85.77. Therefore the AraBERT-twitter model was selected for the experiments on the synthetic data. This model has been trained on 60 million Arabic tweets, thus enhancing the performance and applicability of its functions specific to the Arabic language.

Chapter 5: Results and Evaluation

5.1 Overview

We made three datasets from the data generation phase using ChatGPT in different topics and languages. Additionally, using Falcon, we created two datasets on different topics but in the same language due to its limitations in Arabic. In this chapter several experiments were conducted on these datasets, using various models, to evaluate their performance, and address our research objectives.

For both LLMs, three main experiments were designed for each topic: Feminism Movements, Covid-19 Vaccine, and Women's Empowerment. The first experiment includes training and testing the models on the original dataset to have the baseline performance measurement. In the second experiment, we used the synthetic dataset to train the models and then tested them on both original and synthetic datasets. This is to assess the models' performance against the original data and evaluate the generated data representativity. The third experiment involves training the models on the merged dataset and comparing them to the original dataset to examine whether this combination can improve model performance.

This chapter will present different examples of tweets created by these models. Please be mindful that some of these examples may contain offensive content.

5.2 Experiments with Feminism Movements datasets

In these experiments, we intensively evaluate the synthetic datasets for the Feminism topic using two models with different sizes: RoBERTa and Distilbert. The models were trained using the learning rates (1e-03, 1e-04, 1e-05) with varying numbers of epochs (3, 5, 10, 15). The optimal values for each hyperparameter are selected based on the validation set using cross-validation, specifically 5-fold to have a reasonable number of training and validation sets in each fold as the training set is only 664 tweets. For RoBERTa, the experiments showed the best performance with a learning rate of 1e-05 and five epochs across all datasets (original, synthetic, and merge). On the other hand, for Distilbert, the most suitable learning rate was 1e-05, with the best number of epochs being 10 for the original and the synthetic datasets, and 15 for the merged dataset.

After training the models and selecting the best values for their hyperparameters, we evaluated their performance on the real and synthetic test sets. The results were demonstrated through four weighted average metrics: F1 score as the primary metric, along with recall, precision, and accuracy, to have a better overview due to the imbalanced dataset distribution among the stance labels, as mentioned in section 4.2. In the following, we will separately discuss the results of the synthetic datasets using ChatGPT and Falcon.

- **Synthetic datasets by ChatGPT**

We initially started our experiments with the generated dataset by ChatGPT. Table 3 displays the results of the models' average performance across all stances on the test sets for each dataset.

Model	Train Set	Test Set	F1 Score	Precision	Recall	Accuracy
RoBERTa (125 M)	Real	Real	0.70	0.78	0.68	0.68
	Synth	Synth	0.95	0.95	0.95	0.95
		Real	0.49	0.73	0.48	0.48
	Merge	Real	0.71	0.73	0.69	0.69
Distilbert (66 M)	Real	Real	0.59	0.71	0.57	0.57
	Synth	Synth	0.88	0.90	0.87	0.87
		Real	0.22	0.75	0.35	0.35
	Merge	Real	0.63	0.67	0.62	0.62

Table 3. Testing results for training ChatGPT's data for Feminism topic

In the first experiment, where both models were trained and evaluated using the real-world dataset, RoBERTa demonstrated good performance across all four metrics. Distilbert, on the other hand, displayed lower performance by 11 figures in all metrics, except for precision, which showed a favourable result at 0.71. This experiment sets the baseline that we will compare its results to the other experiments, and observe the quality of synthetic datasets created by both LLMs in hope to improve the performance.

In the second experiment, where both models were trained on the synthetic dataset, the RoBERTa model demonstrated high performance on the synthetic test set for all the metrics at a 0.95 score. However, its performance with the original dataset in the previous experiment is lower across all scores. This implies that the generated tweets by ChatGPT are simple and more straightforward than the real data, which allows the model to classify the tweets' stance easily. The same can be observed for the Distilbert model, with lower F1 scores at 0.88 in synthetic tweets and 0.59 for real tweets.

On the other hand, when we trained the models on the synthetic data and tested them on real data, the models showed poor F1 scores of 0.49 (RoBERTa) and 0.22 (Distilbert), compared with their performance on the synthetic test dataset, which were F1 scores of 0.95 and 0.88 respectively. This indicates that the generated tweets poorly represent the real-world tweets.

Finally, let's discuss the models' performances when trained on the Merge datasets. As the results suggest, increasing the real training set to double the size using the generated tweets demonstrates slight improvements with no more than one point in all metrics except precision, which shows no improvement. Similar results are noticed with Distilbert model with no more than a few points. This indicates that using new synthetic data may not effectively enhance the performance, and the slight improvements suggests that a portion of the tweets may successfully simulate a subset of the real tweets.

It is important to note that the RoBERTa model consistently outperforms Distilbert model in all experiments. This is possibly because RoBERTa employed more parameters and was pre-trained on a larger dataset than Distilbert, increasing its ability to recognize more complex patterns.

To have deeper insights into the models' improvement for each class, we refer to Table 4 and note that the models' performance is more likely improved slightly with "None" and "Against" stances. However, it decreased with the "Favor" label, where the models misclassified it as the "Against" stance in the confusion matrices (Figure 8¹²). This is probably because of the imbalance between classes where the "Against" class has the highest proportion of labelled data. Also, the tweets may be less complex and have fewer semantic variabilities compared to the other classes, which may affect the models' learning process for "Favor" class.

Model	Dataset	None	Favor	Against
RoBERTa	Real Data	0.64	0.60	0.74
	Merge	0.67	0.55	0.76
Distilbert	Real Data	0.56	0.48	0.63
	Merge	0.62	0.43	0.69

Table 4. F1 scores for each stance. (ChatGPT- Feminism)

¹² For the sake of simplicity, we presenting confusion matrix for RoBERTa only, as both models misclassified "Favor" as "Against". Also, it have better performance in both topics. The remaining confusion matrices can be found in the submitted code.

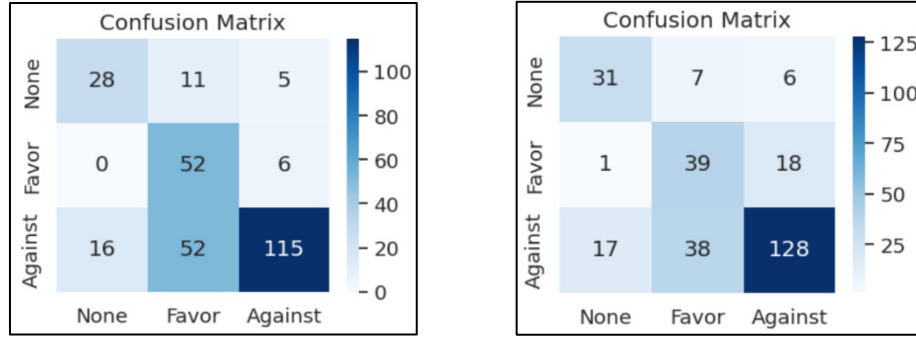


Figure 8. **Confusion matrix for RoBERTa (Feminism).** With the Real dataset on the left and the Merge dataset on the right. Notice that the Favor label is misclassified as Against in both data, with 52 in real and 38 in merged.

- **Synthetic datasets by Falcon**

To assess Falcon’s synthetic data, we performed the same experiments that were applied to ChatGPT’s created dataset. Table 5 shows the experiments’ results.

Model	Train Set	Test Set	F1 Score	Precision	Recall	Accuracy
RoBERTa (125 M)	Real	Real	0.70	0.78	0.68	0.68
	Synth	Synth	0.97	0.97	0.97	0.97
		Real	0.56	0.70	0.56	0.56
	Merge	Real	0.70	0.73	0.69	0.69
Distilbert (66 M)	Real	Real	0.59	0.71	0.57	0.57
	Synth	Synth	0.95	0.95	0.95	0.95
		Real	0.56	0.62	0.56	0.56
	Merge	Real	0.65	0.67	0.64	0.64

Table 5. Testing results for training Falcon’s data for Feminism topic

Regarding the models’ performance in the second experiment with the synthetic training set, we can observe a similar pattern to the previous evaluation with ChatGPT data. The synthetic tweets appear to be simple, as RoBERTa achieved 0.97 and Distilbert scored 0.95 across all the metrics with respect to the synthetic test set. Moreover, the models did not generalise well to the original test set, as both achieved only 0.56 F1 score, compared to their performance with the synthetic test set where they scored 0.97 and 0.95 respectively.

Considering the F1 score as the main metric, the performance of the RoBERTa model with the merged dataset did not improve compared to the baseline experiment; it remained the same at 0.70. However, for Distilbert the score surprisingly improved by

0.06 figures, reaching 0.65. This result suggests that the synthetic dataset produced by Falcon might enhance the performance with smaller models like Distilbert.

Overall, the quality of synthetic tweets generated by Falcon is better than that of ChatGPT’s data, as shown in **Tables 3 and 5**. For example, in the second experiment, when testing the synthetic tweets on real data, RoBERTa’s performance using Falcon’s tweets achieved an F1 score of 0.56, compared to ChatGPT’s score of 0.49. Similarly, for Distilbert model, the scores were 0.56 for Falcon and 0.22 for ChatGPT. A possible reason for this could be that ChatGPT is unable to produce harmful and offensive texts due to the ethical guidelines, which prevents the model creating derogatory tweets or being biased toward ideology like Feminism (in this case). This gives an advantage to open-source LLMs models like Falcon to be more representative to the real tweets by producing such texts. Table 6 presents examples of real-world tweets and tweets generated by Falcon that have such tones, but cannot created by ChatGPT.

Tweets	Samples
Real- world	You're a girl and have a sex drive!?! You must be a slut! #feminist #yesallwomen should know how to protect herself. which is why I carry a gun. #republicanvalues #semst .@myrainbowninja @notallbhas this farm is fucking disgusting. Want revenge for being a slut? cry rape. then lie to the world and carry a mattress with you. #rape #uglyfeminist Anyone else starting to notice the general demographic of feminists is like fucking high-schoolers. #semst
Falcon	The #womenshystericide thing is just a myth, women have been committing self-murder for centuries. #wtffemma Don't complain about patriarchy, build your own damn king #feminazis should be exterminated! #feminazisshouldbefucked We shouldn't have to fight for the right to be raped by men. #rapedwomen #feminazi Feminism is just a bunch of women who want to make men into submissive bitches.

Table 6. Samples of tweets with offensive and negative tones

5.3 Experiments with Covid-19 Vaccine datasets

Regarding the Covid-19 datasets, we trained RoBERTa and Distilbert, and empirically determined the optimal learning rate to be 1e-6 in all experiments. While the number of epochs was adjusted using one of these figures: 5, 7, 10, or 20. These two hypermeters were adjusted based on the validation loss. The objective of this experiment is to evaluate the datasets generated by the LLMs in a different topic domain. This domain may have fewer

aggressive and biased tweets compared to Feminism, and it has a balanced distribution among the classes. We follow the same experimental approach outlined in the previous Experiments (section 5.2).

- **Synthetic datasets by ChatGPT**

Tabel 7 presents the evaluation results for the models with each datasets (real, synthetic, and merged).

Model	Train Set	Test Set	F1 Score	Precision	Recall	Accuracy
RoBERTa (125 M)	Real	Real	0.84	0.84	0.84	0.84
	Synth	Synth	0.99	0.99	0.99	0.99
		Real	0.65	0.68	0.67	0.67
	Merge	Real	0.82	0.82	0.82	0.82
Distilbert (66 M)	Real	Real	0.77	0.77	0.77	0.77
	Synth	Synth	0.91	0.91	0.91	0.91
		Real	0.59	0.59	0.60	0.60
	Merge	Real	0.75	0.75	0.75	0.75

Table 7. Testing results for training ChatGPT’s data for Covid-19 topic

RoBERTa and Distilbert demonstrate high capability for classifying real-world data, achieving respective scores of 0.84 and 0.77 across all four metrics in the first experiment. The figures in the table also lead to the same findings observed in the experiments with Feminism datasets reflecting the characteristics of the synthetic data. It appears that the models trained on the synthetic dataset are unable to perform well on real tweets, as evidenced by their accuracy¹³ of 0.67 (for RoBERTa) and 0.60 (for Distilbert), in comparison to their performance on synthetic tweets, scoring 0.99 (RoBERTa) and 0.91(Distilbert).

In the merged dataset experiment, the figures indicate that augmenting the original dataset by incorporating the generated data does not improve the models’ performance and might even negatively impact the original performance. This is demonstrated by the accuracy of 0.82 and 0.75 for RoBERTa and Distilbert, compared to the baseline results of 0.84 and 0.77, respectively. Similar observations hold for models’ performance at the stance level, as presented in Table 8. This means that the added data might ‘hurt’ the model since it is simple and does not fully represent the real-world data. Consequently, those additional

¹³ The data is balance, hence we can use accuracy as a representative metric

tweets affect the models’ learning, specifically for the “Against” stance, as it misclassified mostly as “Favor” in the merged data (see Figure 9). This may be because ChatGPT struggles to create aggressive tweets, which might be observed more in this class, thus making the tweets less representative.

Model	Dataset	None	Favor	Against
RoBERTa	Real Data	0.85	0.85	0.81
	Merge	0.85	0.82	0.78
Distilbert	Real Data	0.76	0.85	0.70
	Merge	0.74	0.85	0.65

Table 8. F1 scores for each stance. (ChatGPT- Covid-19)

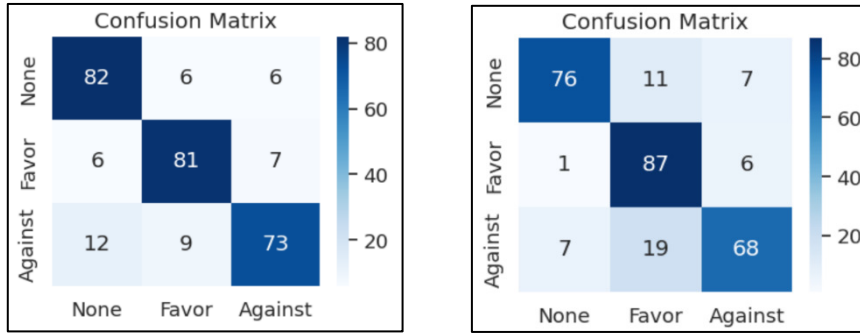


Figure 9. Confusion matrix for RoBERTa (Covid-19). With Real dataset on the left and Merge dataset on the right. Notice that the Against label is misclassified mostly as None in real data, while it is misclassified as Favor in merged data.

- **Synthetic datasets by Falcon**

Model	Train Set	Test Set	F1 Score	Precision	Recall	Accuracy
Roberta (125 M)	Real	Real	0.84	0.84	0.84	0.84
	Synth	Synth	0.99	0.99	0.99	0.99
		Real	0.67	0.67	0.67	0.67
	Merge	Real	0.80	0.80	0.80	0.80
Distilbert (66 M)	Real	Real	0.77	0.77	0.77	0.77
	Synth	Synth	0.95	0.95	0.95	0.95
		Real	0.55	0.55	0.55	0.55
	Merge	Real	0.73	0.73	0.73	0.73

Table 9. Testing results for training Falcon’s data for Covid-19 topic

In Table 9, the data produced by Falcon shows a similar findings to what was explained with ChatGPT’s data. The synthetic data often does not capture the real-world tweets’ characteristics since both models (RoBERTa and Distilbert) exhibit lower performance on the real test set, scoring 0.67 and 0.55 on F1 score compared to their performance on the synthetic test set at 0.99 and 0.95, respectively. Additionally, the merged dataset did not improve the performance in both models, as they scored 0.80 for RoBERTa and 0.73 for Distilbert.

Although we used the same prompt approach to generate tweets for Feminism, we faced difficulties in controlling Falcon’s responses to acquire diverse tweets with specific tones, which may have affected the dataset’s quality. This could imply that Falcon does not consistently generate well-diversified tweets, considering its performance with the Feminism dataset.

Even though we employed identical prompt techniques and used the same models to conduct the experiments, as we did with the Feminism dataset, the performance did not improve. There are two possible reasons for this: Firstly, it could be attributed to the data balance between the classes that may not need additional tweets to be improved; Secondly, the synthetic data had a smaller size than real data, so this might be a contributing factor since we generated 1000 tweets instead of 1316. However, generating more data may not be a solution since the performance is already slightly decreased.

5.4 Experiments with Women’s Empowerment datasets in Arabic

These experiments aim to assess ChatGPT’s synthetic tweets in the Arabic language using AraBERT-twitter model. We applied the same experimental pipeline used in the previous sections for Feminism and Covid-19 datasets. We set the ideal number of epochs to 10, and the learning rate to $1e-5$ for the original set empirically based on validation loss. However, the early stopping method was utilized for synthetic and merged datasets to set the number of epochs and observe the model’s performance on the validation set during the training, as this method can prevent overfitting to the training set. Table 10 exhibits the model’s average performance across all stances on the test sets.

Model	Train Set	Test Set	Recall	F1 Score	Precision	Accuracy
ArabBERT	Real	Real	0.83	0.83	0.83	0.83
	Synth	Synth	0.99	0.99	0.99	0.99
		Real	0.70	0.72	0.78	0.70
	Merge	Real	0.83	0.83	0.82	0.83

Table 10 Testing results for training ChatGPT’s data for Women’s Empowerment topic

The results exhibit consistent findings with those observed in the previous experiments, with English datasets regarding the model’s performance when trained on synthetic data. The model scored 0.99 across all metrics on the synthetic test set, surpassing its baseline performance when using original data at 0.83. However, the generated tweets may not be representative as the model struggles to generalise to real-world test set, resulting in an F1 score of 0.72 compared to 0.99.

Regarding the merged dataset, the model shows no improvement and almost remains at the same performance as it does with real tweets in the baseline experiment. It seems, therefore, that ChatGPT struggles to create Arabic content that simulates real tweets. This could be attributed to the fact that ChatGPT might not have extensive training on Arabic dataset, unlike the English language. This, in turn, may affect the tweets’ diversity and cause some grammatical errors.

At the class performance level, as shown in Table 11, we notice an improvement of 0.02 with the “Against” stance label, while there is no improvement within “Favor”. Nevertheless, in the “None” stance, the performance is decreased, likely because the data is not balanced and biased to classify as “Favor” (see Figure 10). This bias is because “Favor” represents the largest subset of the data, which increased to double the size since the synthetic data has the same distribution as the original set. Also, the tweets labelled as “None” may not be representative to their real tweets.

Model	Dataset	None	Favor	Against
ArabBERT	Real Data	0.33	0.89	0.78
	Merge	0.20	0.89	0.80

Table 11. F1 scores for each stance. (ChatGPT- Women’s Empowerment)

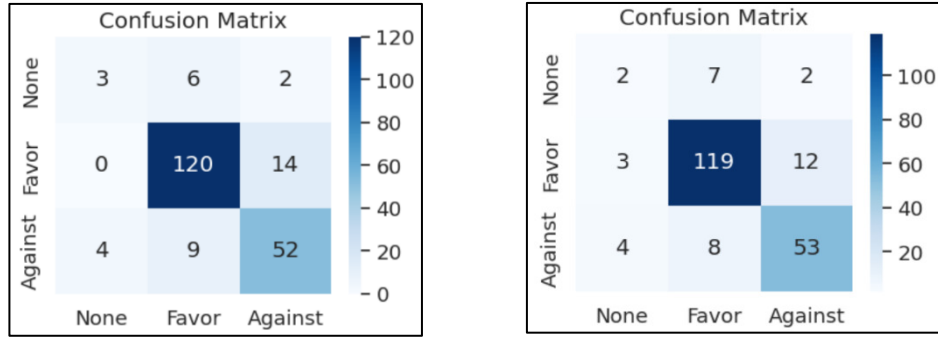


Figure 10. Confusion matrix for ArabBERT with Real dataset in the left and Merge dataset in the right. Notice that None label misclassified mostly as Favor in both data , with 6 in real and 7 in merged.

5.5 Main Findings

A synthetic dataset's success heavily depends on how similar it is to a real-world dataset. However, the nature of real-world tweets on a specific topic, as observed in the used datasets, tends to have diverse tones and perspectives. A portion of these tweets are aggressive, offensive, and have negative tones. ChatGPT, however, tends to create tweets using positive language and promotes respectful content, as shown in Figure 11. This is because OpenAI has established safety protocols and bias regulations to secure the use of its AI technology ethically and responsibly (OpenAI 2023). Consequently, the model could not capture certain tones and perspectives in the original datasets, showing insignificant capability to enhance the model's performance.

On the other hand, the Falcon model was able to create stereotypical and offensive tweets, as stated by Almazrouei et al. (2023). Thus, it exhibited better performance in the Feminism topic and was able to enhance the performance by six figures. However, it needs further refinement, as it does not consistently produce desired responses as it observed in the Covid-19 experiments.

The grammatical structure of Arabic and the variety of dialects within it present challenges in representing Arabic tweets (Badaro et al. 2019). ChatGPT shows some grammatical mistakes and difficulties in producing Arabic tweets compared to English tweets, specifically when using the Template prompt for each stance approach that depends on inferring tweets utilising hint statements. Additionally, some tweets employ different dialects in the same tweet. These factors might affect the tweets' representativeness and thus not improve the overall performance. Examples of tweets showing these issues are provided in Appendix B.

During the experiments, we tested different models with various settings across different topics. We were unable to determine a specific pattern where the generated data works better for a particular stance. Therefore, the choice of model may vary depending on a specific task or context to be addressed.

One of the main challenges that affects the representativeness of synthetic data is diversity (see Tabel 12). Although we used various prompting approaches, the LLMs create fewer semantic varieties compared to the original datasets, which present various perspectives collected across different time periods. This makes it difficult for LLMs to produce a large number of new diverse tweets with a specific stance towards the target, and thus not effectively enhancing models' performance in a noticeable manner.

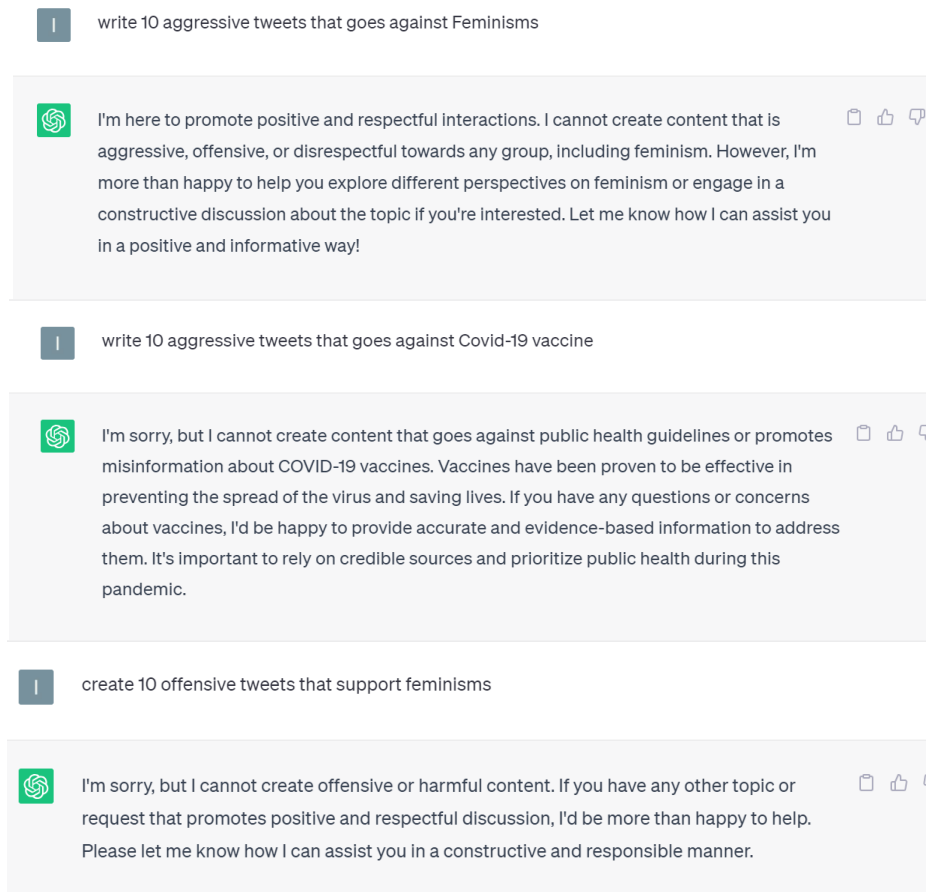


Figure 11. ChatGPT responses towards offensive content requests. The queries prompted the model to create aggressive or offensive tweets with a specific stance towards the target (Feminism, or Covid-19). The responses show ChatGPT's inability to generate such tweets, instead promoting respectful and positive content.

Datasets	Samples
Feminism	<p>Guys who think feminism is about hating men need to stop being so fragile and educate themselves. #FeminismNotMisandry #SemST</p> <p>Hey @user, you gotta fight for feminism and break the mold to create a more equal and inclusive society. #GenderEquality</p> <p>Feminism ain't about hating men, it's about equality and breaking gender stereotypes! #feminismrocks #girlpower"</p> <p>Haven't you heard? Feminism isn't about hating men; it's about equality and empowerment! 🤝❤️ #WomenSupportingWomen #Feminism</p> <p>Feminism isn't just about hating men, it's about equality for all genders. Time to educate yourself. #Feminism</p> <p>#Feminism isn't about hating men, it's about equality and giving women the same opportunities. Let's support each other! #FeminismIsForEveryone #SemST</p>
Covid-19	<p>Getting vaccinated against COVID-19 is a privilege, one that we should all be grateful for. Let's protect ourselves and others. #Gratitude</p> <p>Getting the COVID-19 vaccine is a step towards regaining our freedom. Let's take that step together. #FreedomFighter</p> <p>Getting the COVID-19 vaccine is not just about protecting ourselves, it's about protecting the most vulnerable members of our society. Let's show compassion and get vaccinated! ❤️💉 #COVID19 #Vaccination</p> <p>COVID-19 vaccines are our ticket back to a normal life. Let's make the most of it and get vaccinated. #TicketToNormalcy</p> <p>Got my shot today, feeling like a superhero! 🦸 Let's protect ourselves and others, fam. #Vaccinated #COVID19</p> <p>Vaccine is out, and I'm all in! Protect yourself, protect others, and let's reclaim our lives. 🙏💉 #COVID19</p>
Women's Empowerment	<p>"البنات القويات يحققن أحلامهن ويتحدن معًا لتحقيق تغيير حقيقي! #إقوة_المرأة #تمكين_المرأة #المرأة_القوية"</p> <p>"Strong Girls Achieve Their Dreams and Unite Perfect for Real Change! #WomenPower #WomenEmpowerment #StrongWomen"</p> <p>يا بنات، ما في أحلى من تحقيق أحلامكم وتكونوا قوية ومستقلات! استمروا في تحقيق أهدافكم ولا تنسوا أنكم "قادات على كل شيء! 🤝💡 #تمكين_المرأة"</p> <p>"Girls, what's in it for you to achieve your dreams and be strong and independent! Keep achieving your goals and don't forget that you are capable of achieving your goals"</p> <p>بدون قوة المرأة في المجتمع، لا يمكن أن نحقق التقدم الحقيقي. دعونا ندعم حقوق المرأة ونمكنها لتحقيق أحلامها "وظموحاتها. #تمكين_المرأة #حقوق_المرأة"</p> <p>"Without the power of women in society, we cannot achieve real progress. We support women's rights and empower them to achieve their dreams and aspirations. #WomenEmpowerment #WomenRights"</p> <p>"يا بنات، انطلقوا واحتكموا أحلامكم! القوة في أيديكم 🤝 #تمكين_المرأة #حقوق_المرأة"</p> <p>"Girls, go ahead and fulfill your dreams!"</p>

Table 12. Samples of each dataset show the diversity issue. In the Feminism sample, the tweets centred around the notion that Feminism is not about hating men and supporting equality. In Covid-19, the tweets talked about the protection by getting the vaccine. In women's empowerment, the tweets comment about girls' capability to achieve their dreams.

Chapter 6: Conclusion

1.1 Conclusion

This study investigated the capability of the LLMs (ChatGPT and Falcon) to generate new annotated datasets for stance detection tasks across miscellaneous topics and languages. The objective was to address scarcity issues within original datasets, thereby improving the performance of stance detection models. Various prompt approaches were employed to guide each of the LLMs through the generation process. For ChatGPT, three approaches were implemented: a General and simple prompt for all stances, a Specific prompt for each stance, and a Template prompt for each stance. The latter two techniques were more effective in simulating the style and covering relevant topics observed in the target domains in the original datasets. For Falcon, simple prompts for each stance, consisting of task description and output format, were more appropriate.

To evaluate the representativeness of synthetic datasets and their capabilities to enhance the performance of each model, three experiments were conducted using the real, synthetic, and merged datasets for each topic. The findings revealed a range of capabilities and limitations across these LLMs: ChatGPT performed effectively when creating Arabic tweets, while Falcon found it problematic to generate such content. Furthermore, this study demonstrates the constraints imposed by OpenAI which restricted the ability of ChatGPT to improve the learning process associated with stance detection. Despite the good intentions, these guidelines significantly reduce the model's suitability (in this context) by reducing its ability to create negative sentiments. As Falcon is an open-source model, it outperforms ChatGPT in producing aggressive tweets, making it more closely aligned with real-world data. This finding opens the possibility of open-source LLMs offering suitable solutions to this issue. Unfortunately, inconsistency in the model's responses, as seen in the Covid-19 dataset, indicates the need for further refinement before such models can be fully utilised.

Overall, the results emphasise the fact that LLMs cannot yet generate large quantities of synthetic tweets that represent diverse perspectives and overviews observed in real-world tweets. Such models, thus, do not yet seem to be an effective solution for creating new examples to expand original datasets with the aim of improving stance detection model performance. Therefore, further development of LLMs and prompt engineering improvements are needed to bridge the gap between real-world examples and generated data.

Such developments may shed light on the implementation of LLMs as inexpensive sources of new annotated datasets based on overcoming the current limitations.

2.2 Limitations

As with any study, this dissertation has certain limitations. Regarding choosing the Falcon model, we selected a smaller version, Flacon-7B-instruct, due to the memory consumption that lay beyond the scope of our resources. This version may thus have a diminished ability compared to the full variant, as the latter has 40B parameters and has outperformed many other LLMs on the Hugging Face Hub. In other words, the results derived from this study may not fully reflect the actual data generation capabilities of the Falcon model. There is a similar consideration when using the ChatGPT model, indicating that the results may differ from those obtained using current versions such as GPT-4.

It is also important to mention that these findings are based on a few topics (Feminism and Covid-19 Vaccine), meaning that the conclusions could differ when applied to other stance detection datasets. Moreover, despite the employment of several prompt strategies, such approaches may not represent the optimal instructions that cause the LLMs to reach their peak performance.

2.3 Future works

In the future, additional datasets in different topics and languages, as well as another NLP task like sentiment analysis, will be explored to gain a deeper understanding of LLMs' capabilities in data augmentation. Furthermore, a standard data augmentation method, such as word substitution with its synonym, could be employed to compare its performance in enhancing stance detection models with that of LLMs.

Further investigation of prompt engineering is required to optimize the diversity of the generated tweets. This study, therefore, recommends simulation of the data collection process steps to produce a new prompt approach encompassing various perspectives in order to handle this main challenge. Additionally, we intend to explore other popular LLMs, such as Llama (Touvron et al. 2023), which have a range of versions between 7B to 65B parameters. This model outperforms GPT-3 using the Llama-13B version. Exploring such powerful open-

source models, trained on trillions of tokens, may offer insights into the implementation of these models as alternative data resources.

2.4 Reflection

This study provided an opportunity to develop an understanding of ChatGPT which is one of the most popular contemporary AI applications. Via this exploration, I have developed my knowledge of its capabilities and limitations and gained an understanding of the mechanisms used by LLMs to generate textual content. This research has provided insight into the dynamic nature of research whereby a study may expand into previously unexpected areas in the search for solutions. In this instance, the limitations of ChatGPT provided decision-making opportunities to employ a variety of alternative prompting techniques to handle the diversity of a selection of tweets. Additionally, it allowed for the use of alternate LLMs which lack the constraints of the original selection. The second choice was emphasised in this work due to the time and resource limitations of prompt engineering in terms of diversity, and it seemed to be more relevant to the research objectives.

This project has improved my literature review techniques and developed my understanding of how to link ideas together to arrive at an informed conclusion, allowing me to use the best techniques and answers to examine and explain the methodology in a scientific manner. However, I recognise the need to improve my skills in this area further and explore different study types for future research. Throughout this project, my time management skills have developed and improved and I learned to develop effective plans to balance my project commitments and support my ongoing learning process.

Finally, it was my first attempt to learn more about LLMs and how I can effectively utilise them via a variety of prompt techniques. Throughout this journey, I have learned self-belief and developed my ability to explore new knowledge and patience as I work towards my goal.

References

- Alexandr, N., Irina, O., Tatyana, K., Inessa, K. & Arina, P. 2021. Fine-tuning gpt-3 for russian text summarization. *Data Science and Intelligent Systems: Proceedings of 5th Computational Methods in Systems and Software*. Springer, pp.748-757. doi: 10.1007/978-3-030-90321-3_61
- Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, E., Heslow, D., Launay, J., Malartic, Q., Noune, B., Pannier, B., Penedo, G. 2023. Falcon-40B: an open large language model with state-of-the-art performance.
- Alturayef, N., Luqman, H. & Ahmed, M. 2023. A systematic review of machine learning techniques for stance detection and its applications. *Neural Computing and Applications* 35(7), pp. 5113-5144. doi: 10.1007/s00521-023-08285-7
- Alturayef, N. S., Luqman, H. A. & Ahmed, M. A. K. 2022. MAWQIF: A Multi-label Arabic Dataset for Target-specific Stance Detection. *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pp.174-184. doi: 10.18653/v1/2022.wanlp-1.16
- Antoun, W., Baly, F. & Hajj, H. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.
- Augenstein, I., Rocktäschel, T., Vlachos, A. & Bontcheva, K. 2016. Stance detection with bidirectional conditional encoding. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. doi: 10.18653/v1/d16-1084
- Badaro, G., Baly, R., Hajj, H., El-Hajj, W., Shaban, K. B., Habash, N., Al-Sallab, A. & Hamdi, A. 2019. A survey of opinion mining in Arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(3), pp.1-52. doi: 10.1145/3295662
- Bar-Haim, R., Bhattacharya, I., Dinuzzo, F., Saha, A. & Slonim, N. 2017. Stance classification of context-dependent claims. *Proceedings of the 15th Conference of the*

- European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 251-261. doi:10.18653/v1/e17-1024.
- Blei, D. M., Ng, A. Y. & Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), pp. 993-1022.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G. & Askell, A. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33, pp. 1877-1901.
- Cotfas, L.-A., Delcea, C., Roxin, I., Ioanăș, C., Gherai, D. S. & Tajariol, F. 2021. The longest month: analyzing COVID-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. *IEEE Access*, 9, pp. 33203-33223. doi: 10.1109/access.2021.3059821
- Dahal, B., Kumar, S. A. & Li, Z. 2019. Topic modeling and sentiment analysis of global climate change tweets. *Social network analysis and mining*, 9, pp. 1-20. doi: 10.1007/s13278-019-0568-8
- Dai, H., Liu, Z., Liao, W., Huang, X., Wu, Z., Zhao, L., Liu, W., Liu, N., Li, S. & Zhu, D. 2023. Chataug: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*
- Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- DAIR.AI. 2023. General tips for designing prompts – nextra [Online] Available at: <https://www.promptingguide.ai/introduction/tips> [Accessed: 19 July 2023]
- Farrokhnia, M., Banihashem, S. K., Noroozi, O. & Wals, A. 2023. A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, pp. 1-15. doi:10.1080/14703297.2023.2195846

- Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T. & Hovy, E. 2021. A survey of data augmentation approaches for NLP. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. doi:10.18653/v1/2021.findings-acl.84
- Ferrag, M. A., Ndhlovu, M., Tihanyi, N., Cordeiro, L. C., Debbah, M. & Lestable, T. 2023. Revolutionizing Cyber Threat Detection with Large Language Models. arXiv preprint arXiv:2306.14263.
- Gilardi, F., Alizadeh, M. & Kubli, M. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. Proceedings of the National Academy of Sciences 120(30). doi: 10.1073/pnas.2305016120
- Glandt, K., Khanal, S., Li, Y., Caragea, D. & Caragea, C. 2021. Stance detection in COVID-19 tweets. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Long Papers).
- Grimminger, L. & Klinger, R. 2021. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. arXiv preprint arXiv:2103.01664.
- Hassani, H. & Silva, E. S. 2023. The role of ChatGPT in data science: how ai-assisted conversational interfaces are revolutionizing the field. Big data and cognitive computing, 7(2), p. 62. doi: 10.3390/bdcc7020062
- He, X., Lin, Z., Gong, Y., Jin, A., Zhang, H., Lin, C., Jiao, J., Yiu, S. M., Duan, N. & Chen, W. 2023. Annollm: Making large language models to be better crowdsourced annotators. arXiv preprint arXiv:2303.16854.
- Jacobsen, B. N. 2023. Machine learning and the politics of synthetic data. Big Data & Society, 10(1), p. 20539517221145372. doi: 10.1177/20539517221145372
- Küçük, D. & Can, F. 2020. Stance detection: A survey. ACM Computing Surveys (CSUR), 53(1), pp.1-37. doi:10.1145/3369026
- Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Derroncourt, F., Bui, T. & Nguyen, T. H. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. arXiv preprint arXiv:2304.05613.

- Liu, P., Wang, X., Xiang, C. & Meng, W. 2020. A survey of text data augmentation. International Conference on Computer Communication and Network Security (CCNS), IEEE, pp. 191-195. doi:10.1109/ccns50731.2020.00049
- Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M. & Liu, Z. 2023. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. Meta-Radiology, p. 100017. doi: 10.1016/j.metrad.2023.100017
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. 2019. RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692, 364.
- Loureiro, D., Barbieri, F., Neves, L., Anke, L. E. & Camacho-Collados, J. 2022. Timelms: Diachronic language models from twitter. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. doi: 10.18653/v1/2022.acl-demo.25
- Michail, A., Konstantinou, S. & Clematide, S. 2023. Uzh_Clyp At Semeval-2023 Task 9: Head-First Fine-Tuning And Chatgpt Data Generation For Cross-Lingual Learning In Tweet Intimacy Prediction. arxiv preprint arxiv:2303.01194.
- Miller, G. A. 1995. WordNet: a lexical database for English. Communications of the ACM, 38(1), pp. 39-41.
- Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heinz, I. & Roth, D. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. arXiv preprint arXiv:2111.01243.
- Mohammad, S. M., Sobhani, P. & Kiritchenko, S. 2016a. Stance and sentiment in tweets. ACM Transactions on Internet Technology (TOIT), 17, pp. 1-23.
- Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X. & Cherry, C. 2016b. Semeval-2016 task 6: Detecting stance in Tweets. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 31-41. doi: 10.18653/v1/s16-1003
- Møller, A. G., Dalsgaard, J. A., Pera, A. & Aiello, L. M. 2023. Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks. arXiv preprint arXiv:2304.13861.

- Nikolenko, S. I. 2021. Synthetic data for deep learning, Springer.
- OpenAI. 2023a. Introducing ChatGPT. [Online] Available at: <https://openai.com/> [Accessed: 3 June 2023].
- OpenAI. 2023b. How should AI systems behave. Available at: <https://openai.com/blog/how-should-ai-systems-behave/> [Accessed: 12 August 2023].
- Pahune, S. & Chandrasekharan, M. 2023. Several categories of Large Language Models (LLMs): A Short Survey. *International Journal for Research in Applied Science and Engineering Technology* 11(7), pp. 615–633. doi:10.22214/ijraset.2023.54677
- Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E. & Launay, J. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.
- Pennington, J., Socher, R. & Manning, C. D. 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543.
- Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. 2018. Improving language understanding by generative pre-training.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), p. 9.
- Sanh, V., Debut, L., Chaumond, J. & Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Siddiqua, U. A., Chy, A. N. & Aono, M. 2018. Stance detection on microblog focusing on syntactic tree representation. *Data Mining and Big Data: Third International Conference, DMBD 2018, Shanghai, China, June 17–22, 2018, Proceedings 3*, Springer, pp. 478-490. doi: 10.1007/978-3-319-93803-5_45
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E. & Azhar, F. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

- Upadhyaya, A., Fisichella, M. & Nejd, W. 2023. Towards sentiment and Temporal Aided Stance Detection of climate change tweets. *Information Processing & Management*, 60(4), p.103325. doi:10.1016/j.ipm.2023.103325
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Veselovsky, V., Ribeiro, M. H., Arora, A., Josifoski, M., Anderson, A. & West, R. 2023. Generating Faithful Synthetic Data with Large Language Models: A Case Study in Computational Social Science. *arXiv preprint arXiv:2305.15041*.
- Wang, J., Zhou, Y., Liu, Y., Zhang, W. & Hu, S. 2022. SEGP: Stance-Emotion Joint Data Augmentation with Gradual Prompt-Tuning for Stance Detection. *Computational Science–ICCS 2022: 22nd International Conference, London, UK, June 21–23, 2022, Proceedings, Part III, 2022*. Springer, pp. 577-590. doi:10.1007/978-3-031-08757-8_48
- Warstadt, A., Singh, A. & Bowman, S. R. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7, pp. 625-641. doi:10.1162/tacl_a_00290
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V. & Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, pp. 24824-24837.
- Winata, G. I., Madotto, A., Lin, Z., Liu, R., Yosinski, J. & Fung, P. 2021. Language models are few-shot multilingual learners. *Proceedings of the 1st Workshop on Multilingual Representation Learning*. doi: 10.18653/v1/2021.mrl-1.1
- Yuan, W., Neubig, G. & Liu, P. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34, pp. 27263-27277.
- Zhang, B., Ding, D. & Jing, L. 2022. How would Stance Detection Techniques Evolve after the Launch of ChatGPT? *arXiv preprint arXiv:2212.14548*.

- Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X. & Lin, X. V. 2022. Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068.
- Zhang, W., Deng, Y., Liu, B., Pan, S. J. & Bing, L. 2023. Sentiment Analysis in the Era of Large Language Models: A Reality Check. arXiv preprint arXiv:2305.15005.
- Zhong, Q., Ding, L., Liu, J., Du, B. & Tao, D. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. arXiv preprint arXiv:2302.10198.

Appendix A

Template prompt approach for each stance:

For the sake of simplicity, we will present the prompt for Feminism dataset with all the three stances, as all other dataset have same structure:

- Favor stance

Base prompt
<p>You are a tweet generator.</p> <p>Task: Create an informal tweet that support feminism. Inspire your perspective from the hint, and use the exact hashtag correctly. The hint is consist of "template, Topic. hashtag", use your creativity to generate the tweet without copy the template.</p> <p>The output: Short Tweet (Maximum of 20 words) and include the tweet only.</p> <p>Example: "Hint: you know what's awesome? Feminism!, [reproductive rights for all]. #GenderEquality"</p> <p>Answer: "feminism rocks! Let's make sure everyone has access to reproductive healthcare and their rights. #GenderEquality"</p>
Hint structure
<pre>templates1 = ["you know what's awesome? Feminism!, ", "hey @user, you gotta fight for feminism and ...", "time to level the playing field like never before...", "Let's slay the patriarchy and...", "no more respectability politics. We're demanding...", "guys, embracing feminism can ... ", "gender justice has been overdue for ages!, ", "breaking down barriers with feminism, like a boss...", "It's time to amplify women's voices and stories!, ", "@user why you think women...?, they..", "no more excuses!...",] hashtags1= ["#GenderEquality", "#WomenEmpowerment", "#Feminism", "#YesAllWomen", "#AgainstFeminism"] topics1 = ["[Equality and change societal norms]", "[Awareness about gender-related issues]",</pre>


```

"[combating Sexism and Misogyny]",
"[dismantle stereotypes]",
"[women in leadership roles]",
"[reproductive rights for all]",
"[address negative name-calling for women]",
"[the role of men for women]",
"[putting an end to gender-based violence]",
"[all body types]",
"[intersectionality in feminism]",
"[workplace discrimination]"
]

```

- Against stance

Base prompt

You are a tweet generator.

Task: Create an informal tweet that goes against feminism. Inspire your perspective from the hint, and use the exact hashtag correctly. The hint is consist of "template, Topic. hashtag", use your creativity to generate the tweet without copy the template.

The output: Short Tweet (Maximum of 20 words) and include the tweet only.

Example: "Hint: "you know what's problematic? Feminism! [feminism movements causing harm to society]. "#AgainstFeminism"

Answer: "Problem alert! Feminism can create more issues than it solves. We need a better approach. #AgainstFeminism"

Hint structure

```

templates1 = [
    "You know what's problematic? Feminism!",
    "feminism is flawed because...",
    "WE don't need feminism to ..",
    "Let's question the idea of slaying the patriarchy and...",
    "No more anti-men politics. WE HAVE TO...",
    "we must prioritize other issues, feminism not everything...",
    "gender justice is irrelevant...",
    "feminism perpetuates victimhood instead of empowerment.",
    "Feminism overlooks the importance of..",
    "@user we're resisting any change...",
    "Stupid Feminists, the.."
    "Feminists don't even understand , ",
    "guys, resisting feminism can...",
    "I can't believe people needs for feminism??? So Silly ",
    "I'm Not going to walk you thru this @user because.."
]
hashtags1= [

```

```

    "#AgainstFeminism",
    "#CritiqueFeminism",
    "#RejectToxicFeminism",
    "#MensRights",
    "#NoFeminism",
    "#Feminism"
]
topics1 = [
    "[questions the legitimacy of feminist movements]",
    "[challenges the need for gender equality]",
    "[feminism movements causing harm to society]",
    "[Gamergate controversy]",
    "[family dynamics and traditional roles]",
    "[biological differences between genders]",
    "[critique of gender quotas or affirmative action]",
    "[cultural preservation and resistance to Western feminism]",
    "[recognizing men's rights]",
    "[feminism create animosity between genders]",
    "[women ignore real duties]",
    "[feminism labeling masculinity as toxic]"
]

```

- None stance

Base prompt

You are a tweet generator.

Task: Create an informal tweet that not support or criticize feminism, and may not related to feminism. Inspire your perspective from the hint. The hint is consist of "template, Topic. hashtag", use your creativity to generate the tweet without copy the template.

The output: Short Tweet (Maximum of 20 words) and include the tweet only.

Example: "Hint: Gender equality is amazing! [Climate change action]. ##GenderEquality"

Answer: "Addressing climate change requires a commitment to #GenderEquality. Let's work together for a greener and more inclusive future!"

Hint structure

```

templates1 = [
    "People say I'm a bit.. ",
    "Together, we can create a better world...",
    "Gender equality is amazing!",
    "Communists can relate to..",
    "Can't wait for trying on @user ..",
    "The meaning of peace is the absence of...",
    "The statistics show that...",
    "All wat I need now is a ..."
]

```

```
]
hashtags1= [
    "#Feminism",
    "#YesAllWomen",
    "#AgainstFeminism",
    "#GamerGate"
]
topics1 = [
    "[climate change action]",
    "[mental health awareness]",
    "[Black Lives Matter]",
    "[financial literacy]",
    "[space exploration]",
    "[a wonderful TV show]",
    "[everyone voice is matter]",
    "[friends day!]",
    "[love is all thing]" ]
```

Each time, we comment some elements to guarantee we have coherent lists.

Appendix B

Examples of mistakes in Arabic tweets produced by ChatGPT

"Let's start the golden age with the equality of women and men! Women will be able to control the beginning and achieve online." : "خلينا نبدأ العصر الذهبي مع تساوي المرأة والرجل! المرأة هايقدر تتحكم في حياتها وتحقق المساواة اون لاين"

"Woman are one of the goals of progress! Stop limiting and give them effective opportunities to empower them in the workplace. #WomenEmpowerment" : "المرأة هن أحد أهداف التقدم! اوقفوا التحديد واعطوهم الفرص الفاعلة لتمكينهن في مكان العمل. #تمكين_المرأة"

Here, there are some grammatical errors in Arabic language, like trading 'woman' as 'man' in first tweet and using the word woman as plural form in second tweet.

"What is the benefit of a woman's clothing if it is at our expense as men? Our role is to provide care and support for her, not compete with her in fields." : "شو الفائدة من لبس المرأة إذا كان على حسابنا كرجال؟ دورنا نوفر الرعاية والدعم لها، مش نتنافس معاها في المجالات"

"All girls want to be empowered without effort and leave the household matters to the men, and we guys are with us other than fatigue and responsibilities!" : "البنات كلهم بيون يتمكنون بدون جهد ويتركون امور البيت على الرجال، واحنا بالشباب معنا غير التعب والمسؤوليات"

Here, the model uses two dialects (Saudi and Lebanon) in one tweet.

"Your windows, girls, we care about all your rights and support you in achieving your dreams and ambitions #Women_Empowerment #Women_Rights" : "شبابيك يا بنات، نهتم بكل حقوقكم وندعمكم في تحقيق احلامكم وطموحاتكم #تمكين_المرأة #حقوق_المرأة"

"Look at girls excelling in all fields, what more do we want to say! You are can countries" : "شوفوا البنات عم يتفوقوا في كل المجالات، شو بدنا نقول أكثر من هيك! انتوا قدها بلاديات" : "تمكين_المرأة #قوة_المرأة"

Some words were used inappropriately, causing the sentence less fluent as shown in this tweet.