National College of Ireland

# The Semi-supervised Approach Using GAN and BERT for News Text Classification

MSc Research Project
Data Analytics

Sourav Prabhakar Bhor
Student ID: x19231741

School of Computing
National College of Ireland

Supervisor: Dr Majid Latifi

**National College of Ireland**
**Project Submission Sheet**
**School of Computing**

| | |
|---|---|
| **Student Name:** | Sourav Prabhakar Bhor |
| **Student ID:** | x19231741 |
| **Programme:** | Data Analytics |
| **Year:** | 2021 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Dr Majid Latifi |
| **Submission Due Date:** | 31/01/2022 |
| **Project Title:** | The Semi-supervised Approach Using GAN and BERT for News Text Classification |
| **Word Count:** | 4692 |
| **Page Count:** | 15 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | |
| **Date:** | 2nd February 2022 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | ☐ |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | ☐ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | ☐ |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# The Semi-supervised Approach Using GAN and BERT for News Text Classification

Sourav Prabhakar Bhor
x19231741

**Abstract**

The transformers for Natural Language Processing (NLP), such as Bidirectional Encoder Representations from Transformers (BERT), are impressively effective with these tasks. To test the effectiveness of this model, benchmarks set for the evaluations were made based on the training data, which is humongous in size. In reality, finding the training data appropriately labelled and neatly organised in separate categories is rare at first but also very expensive because of its higher usability. Because training the model to give better suggestions and near-perfect context understanding would require such data in considerable quantity. The semi-supervised GAN models used in image data processing are very effective. This research focuses on fine-tuning the base of the BERT model with the Generative Adversarial Network(GAN) extension. The research aims to provide better performance where the labelled and unlabelled data are meagre. Thus also enabling the use of unlabelled data for the training purpose. I found that the model outperformed the baseline model, especially when the annotated data is below 20 per cent with the dataset I used for this research work.

## 1 Introduction

Machine Learning models are designed to solve specific issues with certain data types, such as text classification use of text classifiers, Natural Language Processing, and even sentiment analysis is industry standard. The Convolutional Neural Network (CNN) is used widely for videos and images.

Today, almost everyone in this digital world consumes news through the Internet. News networks like Google, Yahoo, MSN, Baidu and many more scan the web for news articles. Currently, the entire article is being scanned for context, and it is resource-intensive. The Internet is full of original as well as copied content. The drawback of scanning every single article ultimately means scanning those duplicate contents. Taking this approach ultimately means the wastage of resources. If the same thing can be done based on the shorter yet more effective way, it will save energy for the operation and effective classification in the future. The main reason behind automating this classification is because the news platforms update their feeds to keep their users constantly updated with the latest news(Suleymanov et al.; 2018). Some users would opt-in for precise news alerts that they prefer to get nudged about. From manual classification to machine learning and digital stats in the '90s and now with Artificial Intelligence (AI), the classification tech evolved in every aspect.

Previous research was based on keyword filtering, which considered a specific bag of words found in the distinct possibility that the percentages were highlighted with the assigned category. The main flaw of this technology, if considered in today's scenario, the main flaw of this technology is that many web pages like the news articles that might be covering crime scenes will be flagged as criminal activity category websites instead of highlighting it as a news source(Wu and Hu; 2005). Some technologies used in the past to determine the context are limited to the short description of the news content, which remain useless when determining the exact category of the news article(Singh and Jain; 2021). Consider that the article has only twenty words in the description, which does not cover the news article's entire scenario. This approach will not only classify the news in the wrong category but miss out on the audiences it was intended to target in the first place. The wrong classification of any article will lead to irregular content appearing in wrongfully different categories.

**Motivation:**

The motivation behind this project is to optimise the process of correctly classifying the unannotated data or missing or unlabeled data into specified categories. The research is the capstone project to minimise the time and training data for any new model for the English text News articles based on the headline titles. The research work will help train the model with an even lesser length of the headline and classify it based on its context.

**Research Question:**

How effectively can GAN combined with BERT classify news headlines with meagre labelled data?

**Research Objectives:**

- Study on text classification and state of the art techniques.
- Proposing Research Methodology to answer the research question.
- Text classification using the base BERT architecture to identify its usefulness on the dataset used for this research work.
- Extending the BERT with GAN to classify news effectively and minimise the time for training the model for any new scenario with as little labelled data for training as possible.
- Evaluating the optimisation and effectiveness of the news classification based on the news title with an extension of BERT using GAN and comparing it with the basic BERT model.

**Contribution:**

In this research, the work based on GAN-BERT is used to develop a satisfying accuracy while classifying unannotated data fed to the model. This research is helpful to classify the data which is wrongly labelled or even when data is missing for the input. This research will significantly reduce the resource-intensive tasks such as scanning the entire article for the classification and classifying the content based on the significantly less training data available at hand. This is helpful to train the model faster and put it to use for classification.

**Report Structure:**

This report is divided into multiple sections following the introduction. Section 2 covers the literature review for this field's work done to date. In the section 3, I propose

the research methodology that is used in this research project. Section 4, discusses the project implementation. Then in the later section 5, the output summary and related evaluation strategies are discussed to evaluate this research work and also covers the results and critical analysis conducted on the work done in this project. The final section 6 includes a further discussion on a conclusion of the work done and possible future work that can be done considering this research work as a base.

# 2 Related Work

Deep Learning approaches have grown in popularity in Natural Language Processing (NLP), for example, since they achieve excellent performance by depending on relatively simple input representations (for example, in NLP). Natural Language Processing (NLP) has various dimensions in the actual field.

The report's related work is divided further into two separate subsections wherein the subsection 2.1 out of two, techniques applied on text classification using Natural Language Processing (NLP) are discussed, and in subsection 2.2, discussion of semi-supervised BERTwith Adversarial Learning.

## 2.1 Text classification using Natural Language Processing (NLP)

The material on the Internet is growing every day, offering internet users a wealth of information. Unfortunately, because of several other factors, this scenario could end up causing more harm than benefit to online users. It contains damaging or deceptive information. The toxic contents can be anything from textual to graphical regarding violence, pornographic material, or any other potentially dangerous material. Emotionally traumatic visuals or descriptions harmful to children and adults lead to various web blocking approaches, such as adblocking, which restricts access to these dangerous materials. The approaches include Uniform Resource Locator(URL) block list, DNS level intelligent analysis, and semantic web content analysis. This paper suggests a method for categorising web pages that include adult material, with a BERT accuracy of 67.81 per cent among 32 distinct categories. This paper demonstrates that a BERT model provides more precision than the Sequential and Functional APIs used for text categorisation models (Demirkıran et al.; 2020).

The research work done by Zhang and Yamana (2021) describes several algorithms and models, most often adding the valuable insights buried in different classifiers. Researchers contend that concealed material in class labels improves classification performance. Rather than converting the labelling into numeric data, this paper included the information in the initial formulation without modifying the architectures. Researchers integrated the results of an initial classification algorithm with connectedness obtained from sequencing and keywords set vectorisation. The approach is helpful to make sense of the text context instead of randomly generating the text sequence. A keyword set is a group of words used to convey information in labels. It is often created by classes, although users may also change it. The research findings suggest that the suggested approach significantly improved text categorisation tasks. The approach required have a lot of labelled data which could be impossible in many cases, and our research mainly focused on lesser training data but gave the most accuracy. The lack of derived features complicates text categorisation. If the information system cannot effectively and adequately generalise information categorisation and user interest suggestion, it will undoubtedly

influence platform customers' experience and usage. In this study, the approach is based on text categorisation and personalised recommendations for subscribers. The Wide and Deep-BERT model is built on the Wide and Deep model with the upgraded BERT data preprocessing model.

Furthermore, the associated news text categorisation and recommendations technology process is presented. The Tensorflow deep learning framework is utilised to scientifically validate the innovation, demonstrating the efficacy and productivity of the design methods (Jing and Bailong; 2021). The way TensorFlow was used is used in this research to check the validity of our research's validity. Throughout this research, researchers thoroughly examined multiple Bert-based perfect methods for diverse text categorisation problems. Various kinds of tweaking Bert models, each with a various classification layer, are constructed, and their efficiency is rigorously testified. Deep learning networks such as CNN architectures and Bi-LSTM were used in the perfect models built. Each model's feature extractor was researched to get completely different inputs from Bert's many layers (Mohammadi and Chapon; 2020). Recent approaches for developing word2vec representations successfully captured all right semantic and syntactic patterns using vector computing, but the source of these behaviour patterns has remained unknown. Researchers investigate and explain the model features required for such constants to develop in word representations (Pennington et al.; 2014). News text categorisation is a significant undertaking that easily captures everybody's interest in our everyday lives. This research involves the BERT model in the Transformers architecture. The short and long memory networks of the RNN are evaluated to the identical news text data set. According to experimental data, the categorisation efficiency of the BERT framework is much greater than that of the long and short memory networks (Deping et al.; 2021).

This research is most important to go further with the BERT based framework to go ahead and classify news based on the headlines with the minor data to train and input but with higher accuracy. This paper provides a text analytics (NLP) strategy for automating ad texts published on internet advertising networks. Data collection contains around 21,000 tagged advertisement texts spanning 12 industries. The Bidirectional Encoder Representations from Transformers (BERT) model was employed inside this research (Özdil, Arslan, Taşar, Polat and Ozan; 2021). With the help of this approach, the ad categorisation proved to have a better categorisation of the ad uploaded by the clients automatically. The research conducted by Wang and Song (2019), shows the use of deep learning in text classification, combining it with the features of news content, and proposes a twofold Bi-Gated Recurrent Unit (GRU) Plus attentiveness deep learning method to forecast spikes, with impressive outcomes. The study primarily researches news text categorisation in this paper. It presents a Latent Dirichlet Allocation-based news text categorisation model (LDA). Because the scale of the news content is too large, this study utilises the classification method to minimise the word dimensionality and get attributes. The approach is handy in this research, reducing the headline length to a particular word count. Simultaneously, the study conducts work on the Convolution layers regression technique to address multi-class text issues in our daily lives and use it as a model's classifier (Li et al.; 2016). This study aims to learn the SVM approach for news classification using Indonesian news datasets with various news categories. The quantity of characteristics that impact classification efficiency using SVM is one of the difficulties in text categorisation. The use of Extracted Features as a selecting feature improves the accuracy over no extracted features. The algorithm produced a satisfactory outcome with 99,057 per cent efficiency in the Indonesian news categorisation problem

statement. SVM approach without extracted features improves by 2,8 points from 96,11 per cent (Rizaldy and Santoso; 2017).

## 2.2 Semi-supervised BERT with Adversarial Learning

A pre-training stage of transformer-based designs reflects their inputs. These are trained on large datasets, which consist of thousands of data record entries or even more and these transformers are then fine-tuned over a specific problem statement to get state-of-the-art outcomes in those NLP tasks (Kim; 2014). When thousands of labelled data samples were provided for the last tasks, accuracy skyrocketed to reach new highs. The research demonstrated by Devlin et al. (2019), the classification quality of BERT with less than 200 labelled samples suffers a significant decrease, where multiple categories were provided. However, obtaining tagged data takes a more extended period and is expensive. Adopting semisupervised approaches, such as in this case, to increase generalisation capabilities when limited annotated data sources are available while accumulating unlabeled sourcesYang et al. (2016).

Semi-Supervised GAN is one of the successful semi-supervised approaches(SS-GANs). A "generator" is often trained to create the most identical samples that resemble some input data. This "adversarial" training approach is dependent on a "discriminator," which is instead instructed to determine the generator samplings from actual inputs. SS-GANs are a GAN extension in which the discriminator provides a category to each sample, simultaneously finding whether it was created automatically by the system's generator or not. Therefore, the labelled material is utilised for training the discriminator in SS-GANs, while the untagged and generated samples enhance its internal representations. SS-GANs are successful in image processing: when subjected to a few labelled examples with more untagged ones, they achieve satisfactory results comparable to fully supervised settings.

While reviewing this study, training the BERT with untagged data in a generative adversarial context is extended. The GAN-BERT model specifically enhances the BERT's fine-tuning process from the viewpoint of SS-GAN. A generator generates "bogus" samples similar to the data input string to the model, and as a discriminator, a BERT is employed. In this approach, for the final tasks, researchers have used both BERT's capacity to use unlabeled material to assist the network in generalising and building high-quality representations of input texts. The Kernel-based GAN (KGAN) is used to examine SS-GANs in NLP. With the perspective of SS-GAN, an extension of a Kernel-based Deep Architecture (KDA) is studiedCroce et al. (2017).

Projections of the sentences are embedded into low-dimensional embeddings, corresponding to the inferential space developed using a Semantic Tree Kernel function. GAN's perspective of enhancement of deep architecture for NLP tasks is only considered. Approximating a pre-computed embedding space by a kernel function is worked up by a KGANAnnesi et al. (2014). Multi-layered perceptron quality is enhanced by the SS-GAN utilised in the KDA. Statically obtained by the kernel space approximation, the input representation space is unaffected. All network parameters are considered throughout the training phase with SS-GAN techniques.

Many research works presented here took various approaches from structure and visual level inspection of the data to detect and classify the data based on it. Such approaches were severely impacted when the data was slightly modified or noisy data was included at

the source. Scanning and analysing content frequently to keep the classifiers up to date with the latest data requires a compelling resource. Such approaches may sometimes be unoptimised for text classification and produce inaccurate results. These approaches are not helpful while generalising content based on the context and may not hold much authenticity to categorise the content with more outstanding trust scores. The main advantage of BERT, which is developed and backed by Google, is that it comes with the optimised and pre-trained model, which requires very little training and is not as heavy as training similar models from scratch. It saves time on training and gives access to broader text resources for our trainable data when our model is given input from BERT output.

The GAN model used in this research can create noisy data indistinguishable from the original data. This gives an advantage for the models trained with less labelled data to develop a more robust model to provide better results and authenticity to the problem they are developed for. Due to their ability to detect and duplicate the data as closely as possible, they can be trained even quicker to understand the context of the work. The discriminator in GANs takes loss values from both labelled and unlabelled data into account. When a real example is assigned a wrong category, loss values for labelled data are generated. When a fake example is wrongfully categorised with an actual label, loss values for unlabelled data are generated. These loss values help improve the classification ability of the discriminator and the fake data generation ability of the generator. The main disadvantage of it is that they are tough to train, and the limitations on them are only that they are required to be checked very frequently for their effectiveness. It is tough to train such models, and textual data are very complex to understand, and the models based on these inputs are harder to train.

# 3 Methodology

The research work is helpful to classify text on the minimum amount of data and which can be processed in large batches. For this research, data gathering is carried out from an accessible repository across the globe, e.g., Kaggle, as datasets. The data obtained from the dataset is clean, but some preprocessing is done to balance the dataset. To tackle the issue with the ongoing research where categorisation will occur based on the headline, Natural Language Processing(NLP) is to be used. Transformer-based architectures are more suitable as they are masked language models that are particularly useful for detecting the text's context. Hence, the BERT and the GAN are developed and deployed for research.

## 3.1 Data Collection

I use a news dataset with over 200K entries containing the category, news headline and a short description for each data point [1]. The data has been collected between the year 2012 to 2018 from HuffPost[2]. I have been using this dataset collected and compiled by the independent author and given access for public and research use without any licensing terms but with the citation request. I have provided the properly defined citation for the

[1]https://www.kaggle.com/rmisra/news-category-dataset
[2]https://www .huffpost.com/

author in terms of data sources. I have had the single JSON file downloaded on our local machine.

Table 1: Attribute Summary of dataset

| No. of | Before | After |
|---|---|---|
| Samples | 200,853 | 40,300 |
| Classes | 41 | 31 |
| Empty Headlines | 6 | 0 |
| Missing Headlines | 0 | 0 |
| Samples per class | Min: 1K Median: 3.4K Max: 32.7K | 1300 |
| Words per sample | Min: 0 Median: 10 Max: 44 | Min:1 Median: 10.0 Max: 38 |

## 3.2 Data Preprocessing

I then process the data to remove all empty headlines and combine various synonymous categories. This brings down the count of classes from 41 to 31, and the corresponding class distribution is shown in Figure 1. This will enable the model to distinguish between headline styles better and decrease ambiguity between similar categories.

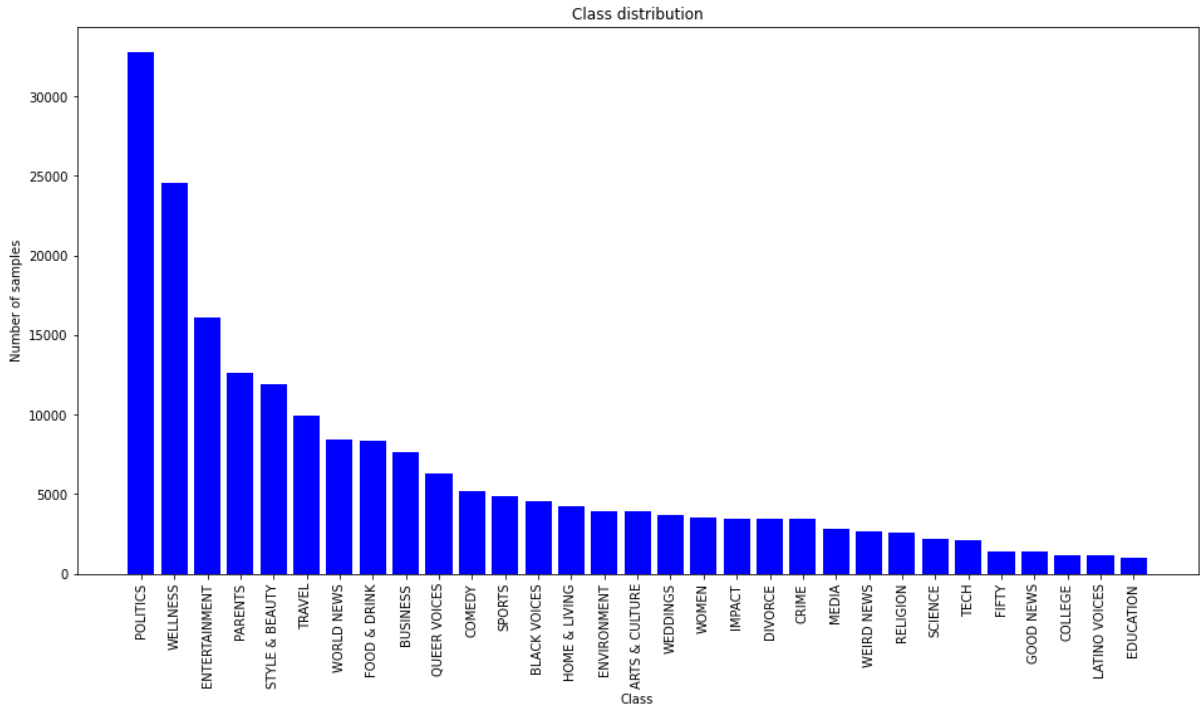


Figure 1: Class distribution of categories after preprocessing

To remove class imbalance, I have collected an equal number of samples from each category and settled at a sample size of 40,300 with 1300 samples for each category. The sample size was chosen while keeping the training time and number of different experiments in mind. The summary of the unprocessed and the processed data can be found in Table 1

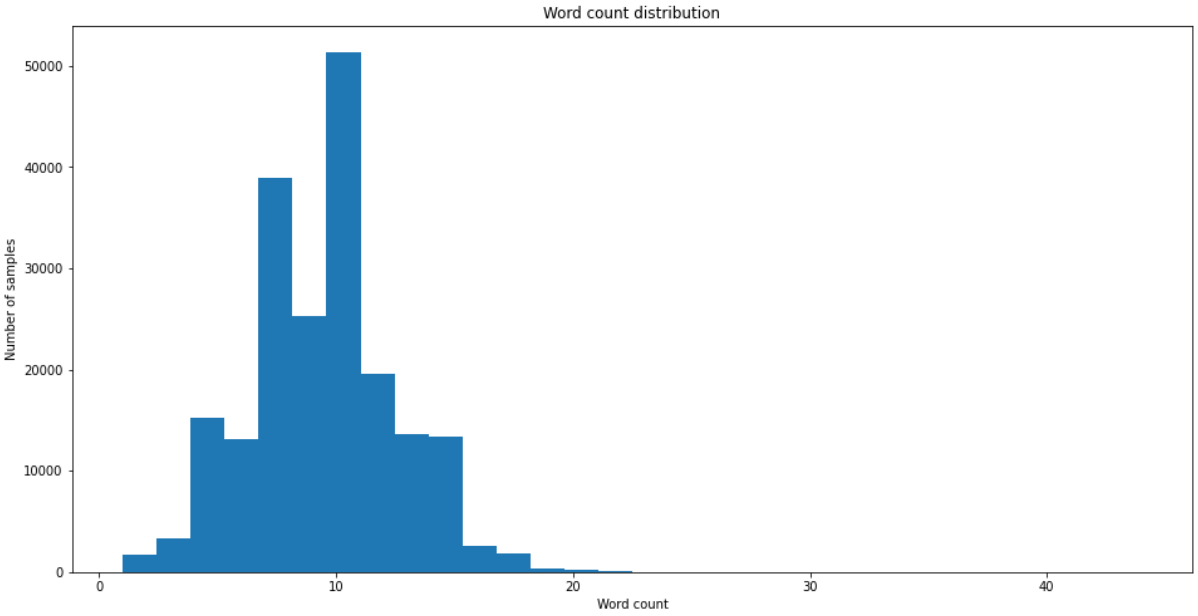


Figure 2: Word Count distribution after preprocessing

The word count distribution can be seen in Figure 2 which indicates that most headline sequence lengths lie between 0 and 22, with a minuscule amount of them breaching our maximum sequence length of 32, which has been defined to strike a balance between training accuracy and time.

## 3.3 Model Development based on GAN and BERT

Various sequence length and batch size restrictions apply when training BERT based models on Graphics Processing Unit (GPU). There is always a trade-off between batch size and sequence lengths when training systems with limited GPU resources. Smaller batch sizes increase training time but allow more extensive sequences as input and vice-versa. The nature of news headlines is meant to be short with higher information entropy, thus negating the usefulness of more extensive sequences as input. The median sequence length is 10 in our dataset, with a maximum length at 44 as shown in Table 1. Thus keeping the sequence length as 32 enables us to try and test different BERT models without any significant accuracy loss.

## 3.4 Comparison & Evaluation

To assess the performance of our model for news classification, I carried out experimentation over various percentages of labelled data. I used both BERT-base-cased[3] and BERT-base-uncased[4] to study the impact of using a cased-based-pre-trained model on news headlines and if there are any significant advantages in anyone.

My developed model begins with significantly less annotated data and gradually increases it. This will be used to study the varying performance of our semi-supervised and how close it can get to a supervised model's performance, more specifically at what

[3] `https://huggingface.co/bert-base-cased`
[4] `https://huggingface.co/bert-base-uncased`

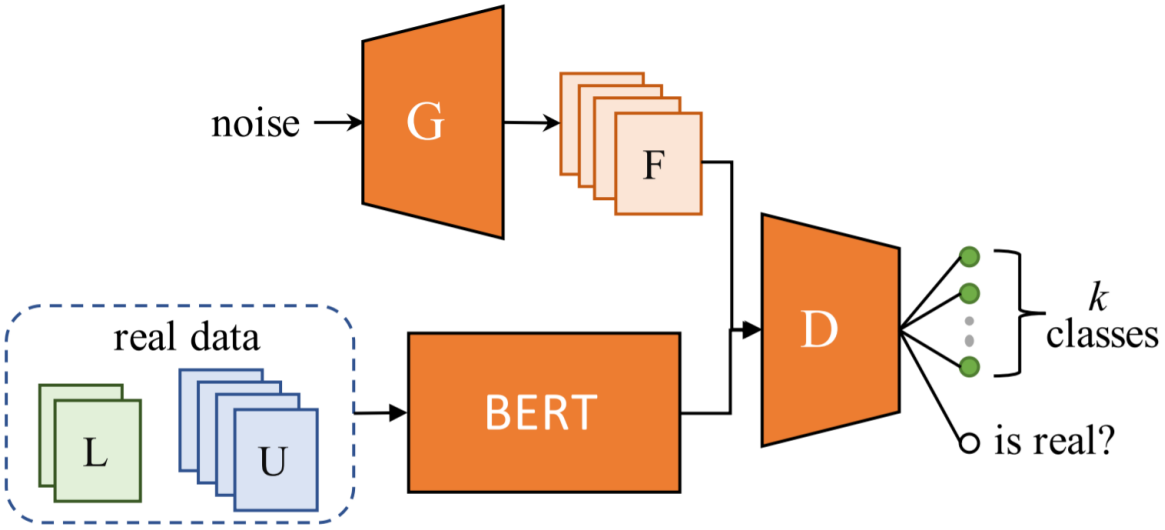


Figure 3: GAN-BERT Model Simple Illustration

amount of annotated data. I use a simple BERT classifier[5] as a baseline supervised model in order to compare with my developed GAN-BERT model. The most commonly used metric, i.e., accuracy, has been used in this research as I have preprocessed the dataset to balance all classes. The test set has been kept constant across all experiments (both supervised and semi-supervised) to get a fair assessment of the model performance.

# 4 Implementation

This research work is based on Natural Language Processing to classify news headlines into pre-defined categories. In this research work, transformers are the new family of neural network architectures proposed in 2017 and adopt the self-attention mechanism for each input item. The transformers use the attention for each input section and compare it with the other sections instead of comparing it with each output section.

Generative Adversarial Network (GAN) extends the regular BERT model where semi-supervised learning is employed. The unannotated data for the training session of this model is a big problem. Significantly few annotated data does not help build a super-powerful model to provide sufficient accuracy for classification of the text for the required purpose. This issue can be resolved using the fine-tuning stage, found in the GAN-BERT model using discriminator-generator. The GAN-BERT model is a multi-headed transformer as shown in figure 4.

The discriminator in the GAN-BERT model, as shown in figure 5, discriminates valid and invalid data, hence the name discriminator. The generator in this model is used to create noisy or fake data and will train the model to detect nearly impossible noisy data from the valid data. The discriminator has classes in which it is supposed to classify the data coming from the generator and actual input data. Both the generator and discriminator plays the roles based on the penalties for each other's mistake. The generator is responsible for generating nearly similar data to the actual data. However, when the discriminator correctly identifies the data and classifies it as false, the generator is penalised,

[5] `https://huggingface.co/docs/transformers/index`

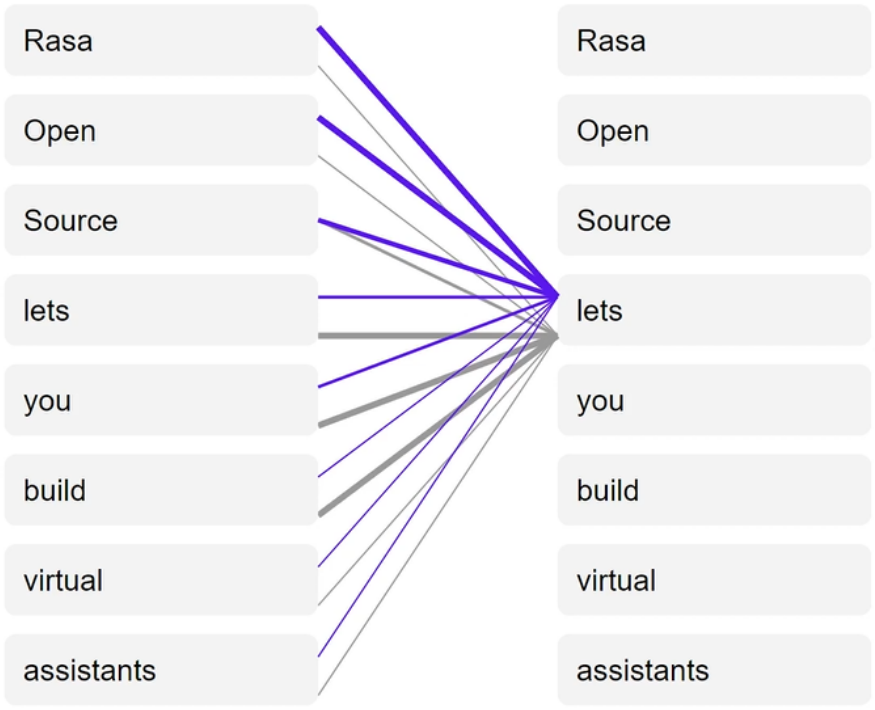


Figure 4: Multiheaded Transformer

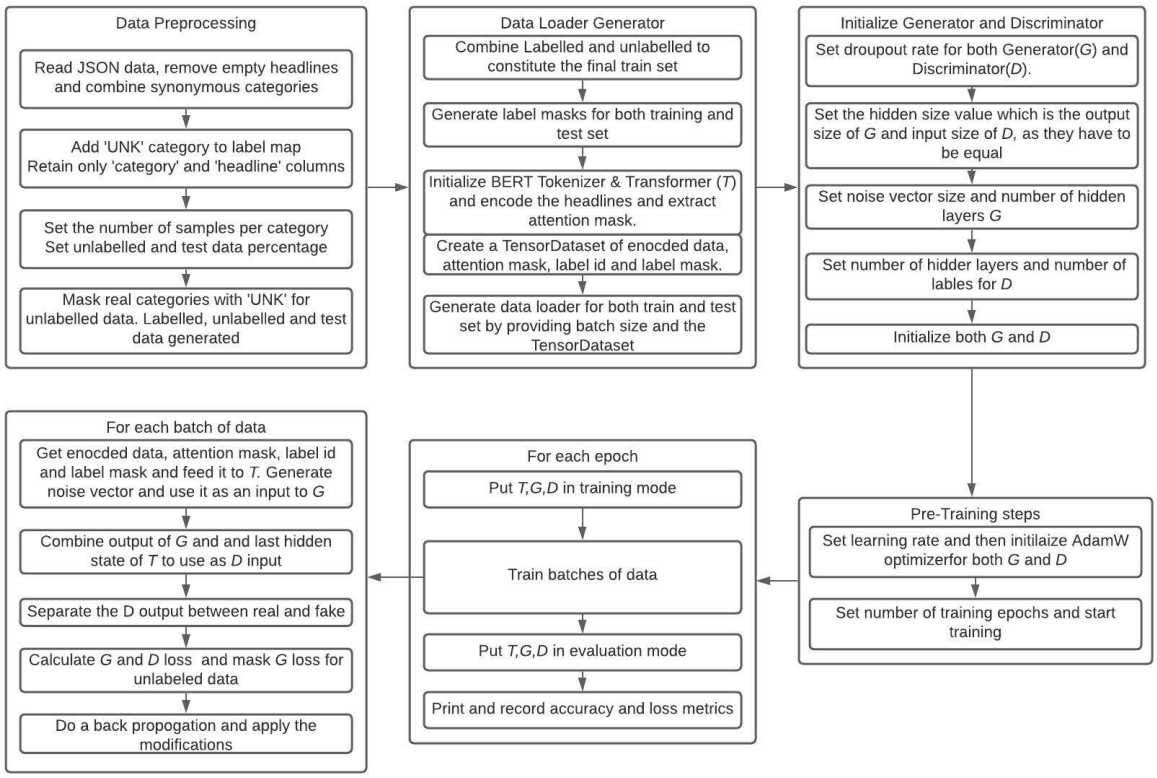


Figure 5: Pseudo Code

and the same goes in the other direction.

This model 5 is then subjected to the input feeds to train appropriately, where the labelled data is first provided, and the classification accuracy is measured. Then in the next phase, data generated by the generator is fed to the discriminator and the ability of the discriminator is tested on how well it classifies the noisy data from the generator. Then after the training, both data are provided to the model at the final stage, which includes labelled and unlabelled data simultaneously, and the accuracy and ability of the GAN-BERT model are tested.

The work on this project is entirely coded in the Python language. Some libraries specific to the project and models developed by me are used. NumPy is required for mathematical operations efficiently. As this research work is primarily focused on text classification, hence I am taking the help of Natural Language Processing, PyTorch is used for it. PyTorch is helpful in this research precisely because of its ability in type-based automatic type differentiation, which is used to evaluate the derivative function of the model set for the research work. TensorFlow is also used to determine specific parameters and validate some claims in this research work. Tensor computing is very resource-intensive, and hence running the operations based on the CPU itself will take longer than usual. The local machine used for this research work has powerful GPU cores to make the model more efficient. PyTorch can run all the tensor computation on those cores for stronger acceleration.

# 5 Evaluation

In this section, I have investigated the model performance over various characteristics and training conditions. I tested our model with varying levels of annotated data and two types of BERT model, i.e., BERT-base-cased and BERT-base-uncased. In BERT uncased, all texts are lower-cased before feeding it into WordPiece tokenisation. While in BERT-uncased, the texts are maintained to contain more information about proper nouns and part-of-speech tagging. Part-of-speech tagging provides more information about sentence structure, which might prove to be more useful in certain conditions like long texts and sentence prediction. Thus, I experiment with both types of BERT models to see if it affects text classification news headlines in our case.

Table 3: Labeled/Unlabeled data count for each annotation setting

| **Annotation %** | **#Labeled Samples** | **#Unlabeled Samples** |
|---|---|---|
| 5 | 1612 | 30644 |
| 7.5 | 2419 | 29837 |
| 10 | 3225 | 29031 |
| 20 | 6451 | 25805 |
| 30 | 9676 | 22580 |
| 40 | 12902 | 19354 |
| 50 | 16128 | 16128 |

The test set has been kept constant across all experiments, containing 8,044 samples. All the categories are balanced in our test set; thus, accuracy will be the best fit for my comparison. The training set contains a total of 32,256 samples. I used various

annotation percentage settings as shown in Table 3. Going with annotation percentage below five was giving inconsistent results with huge standard deviation; thus, those results were not considered and evaluated. As mentioned in Section 3.4, I used a BERT-based simple classifier which is a supervised model, as a baseline to compare it against my semi-supervised model.

## 5.1 GAN-BERT vs Baseline

The classification accuracy of my model is shown in Figure 6 denoted by the orange line while the grey line denotes the baseline model. The plot's x-axis represents the accuracy percentage, while the y-axis shows the amount of unlabeled data.

My model achieves 41.8% accuracy using just 5% of labelled data, while the baseline model achieves 27%. This trend continues till 50% of labelled data, where the accuracy of both of them get very close, and this is because BERT in itself is a compelling pre-trained model. As I increased the annotation percentage, the model gradually tended to perform similarly to a supervised model with a diminishing advantage, as shown by the converging lines.

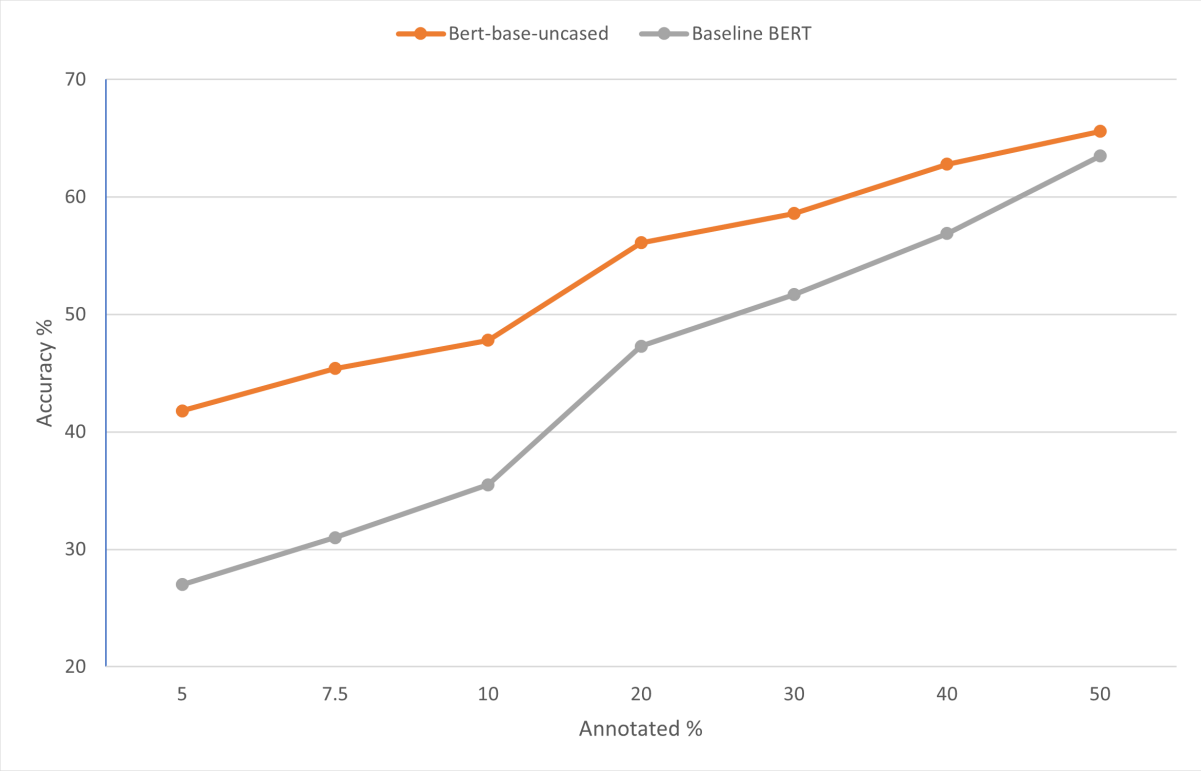


Figure 6: BERT-base-uncased vs Baseline

## 5.2 BERT-base-cased vs BERT-base-uncased

This experiment has been conducted to assess if the BERT-base-cased model can improve the GAN-BERT performance as news headlines contain many Upper-cased words. BERT-base-uncased outperforms the cased model in most NLP use cases, except in some specific cases. Here, I see that the uncased model outperforms the cased model as found in most

cases which indicates that cased words in the headlines do not provide any additional information to GAN-BERT; in turn, it reduced performance.

At 5% and 50% annotated data, the uncased model gives 5.3% and 1.9% more accuracy correspondingly than the uncased model as shown in Figure 7. It still performs more than the baseline model, reaffirming combining the GAN architecture with BERT.

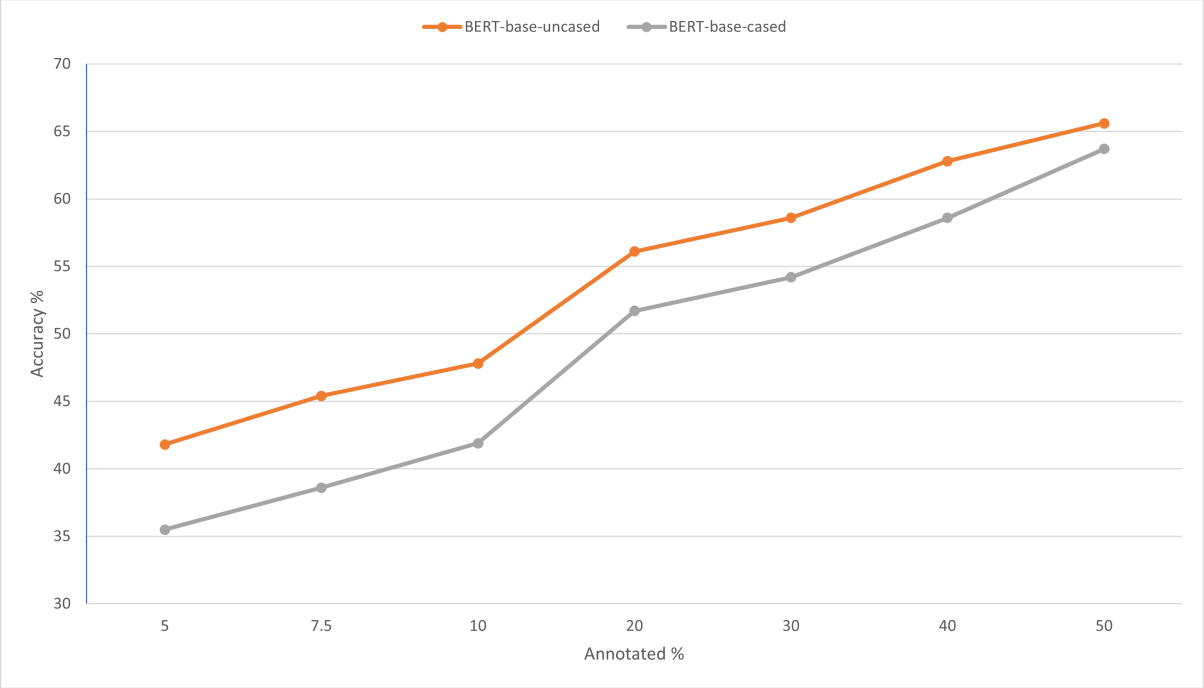


Figure 7: BERT-base-uncased vs BERT-base-cased

# 6 Conclusion and Future Work

Transformer-based architectures are limited and perform poorly when trained, like minimal data or a considerable portion of unlabeled data. During the analysis for this research, experiments show that models built upon minimal labelled data were volatile. Their performances were not reliable enough to be used in real-world applications. Semi-supervised learning based on transformer architectures may prove to be the effective way for such cases. The main motive of this research is to improve the effectiveness of BERT and systematic improvement using generative adversarial networks(GAN) without investing additional computational and memory costs into the existing architecture. The increased robustness of our model can pave the way for further research in the field of semi-supervised learning using transformer like models.

Using a higher annotated sample results in a very similar level of performance, but GAN-BERT always has a slight advantage. The future scope for this project would be to use the way to several transformers extensions, including the adoption of other architectures, such as GPT-2(Qu et al.; 2020), DistilBERT(Mozafari et al.; 2020), RoBERTa(Liu et al.; 2019) and others. The results in this research hint at better semi-supervised performance; thus, different architectures can be selected and fine-tuned for specific text classification cases. Moreover, further investigation can be done to study the potential

impact of BERT pre-training by direct adversarial training. Also, it would help to investigate the produced representations from the generator encodings to gain insight from a linguistic perspective.

# References


Annesi, P., Croce, D. and Basili, R. (2014). Semantic compositionality in tree kernels, p. 1029–1038.
**URL:** *https://doi.org/10.1145/2661829.2661955*

Croce, D., Filice, S., Castellucci, G. and Basili, R. (2017). Deep learning in semantic kernel spaces, pp. 345–354.
**URL:** *https://aclanthology.org/P17-1032*

Demirkıran, F., Çayır, A., Ünal, U. and Dağ, H. (2020). Website category classification using fine-tuned bert language model, pp. 333–336.

Deping, L., Hongjuan, W., Mengyang, L. and Pei, L. (2021). News text classification based on bidirectional encoder representation from transformers, pp. 137–140.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding, pp. 4171–4186.
**URL:** *https://aclanthology.org/N19-1423*

Jing, W. and Bailong, Y. (2021). News text classification and recommendation technology based on wide amp; deep-bert model, pp. 209–216.

Kim, Y. (2014). Convolutional neural networks for sentence classification, pp. 1746–1751.
**URL:** *https://aclanthology.org/D14-1181*

Li, Z., Shang, W. and Yan, M. (2016). News text classification model based on topic model, pp. 1–5.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach, *CoRR* **abs/1907.11692**.
**URL:** *http://arxiv.org/abs/1907.11692*

Mohammadi, S. and Chapon, M. (2020). Investigating the performance of fine-tuned text classification models based-on bert, pp. 1252–1257.

Mozafari, J., Fatemi, A. and Moradi, P. (2020). A method for answer selection using distilbert and important words, pp. 72–76.

Pennington, J., Socher, R. and Manning, C. (2014). Glove: Global vectors for word representation, *EMNLP* **14**: 1532–1543.

Qu, Y., Liu, P., Song, W., Liu, L. and Cheng, M. (2020). A text generation and prediction system: Pre-training on new corpora using bert and gpt-2, pp. 323–326.

Rizaldy, A. and Santoso, H. A. (2017). Performance improvement of support vector machine (svm) with information gain on categorization of indonesian news documents, pp. 227–232.

Singh, A. and Jain, G. (2021). Sentiment analysis of news headlines using simple transformers, pp. 1–6.

Suleymanov, U., Rustamov, S., Zulfugarov, M., Orujov, O., Musayev, N. and Alizade, A. (2018). Empirical study of online news classification using machine learning approaches, pp. 1–6.

Wang, Z. and Song, B. (2019). Research on hot news classification algorithm based on deep learning, pp. 2376–2380.

Wu, O. and Hu, W. (2005). Web sensitive text filtering by combining semantics and statistics, pp. 663–667.

Yang, Z., Cohen, W. W. and Salakhutdinov, R. (2016). Revisiting semi-supervised learning with graph embeddings, p. 40–48.

Zhang, C. and Yamana, H. (2021). Improving text classification using knowledge in labels, pp. 193–197.
ΩÖzdil et al.

Özdil, U., Arslan, B., Taşar, D. E., Polat, G. and Ozan, (2021). Ad text classification with bidirectional encoder representations, pp. 169–173.