National College of Ireland

Evaluating Deep Learning Models for Cross-domain Propaganda Detection

MSc Research Project
MSc Artificial Intelligence

Mohammed Sameer Ahmed
Student ID: X23337257

School of Computing
National College of Ireland

Supervisor: Mr. Taylou Maniganze

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Mohammed Sameer Ahmed
Student ID:	X23337257
Programme:	MSc Artificial Intelligence **Year:** 2024-25
Module:	MSc Research in Computing
Supervisor:	Mr. Taylou Maniganze
Submission Due Date:	13th September 2025
Project Title:	Evaluating Deep Learning Models for Cross-domain Propaganda Detection
Word Count:	4957

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: Mohammed Sameer Ahmed

Date: 11th August 2025

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):	

Evaluating Deep Learning Models for Cross-domain Propaganda Detection

Mohammed Sameer Ahmed
X23337257

Abstract

Propaganda is a major issue in a world generating data and information exponentially. It has a huge impact on our societies and potential to influence people. With advances in technology, the propaganda techniques can be detected and tackled, and previous works have given some excellent results in doing so. While these studies have given applaudable results they are restricted to one single domain of information. Very little work has been done to generalize the propaganda detection and to address this gap this study evaluates performance for four deep learning models (LSTM, BERT, XLNet and GPT-2). Each model is trained on one domain of information and tested to detect propaganda in information form other domains. The results reveal that these models perform detection significantly better in same domain than a different domain. The Accuracy and F1 are high even with limited resources but this drops sharply in a cross-domain setup. Transformer-based models, specifically BERT (Accuracy: 93%, F1: 98% on speech articles) and XLNet (Accuracy: 97%, F1: 98% on news articles), outperform the traditional LSTM (Accuracy: 91%, F1: 92% on speeches) but still struggle to get acceptable scores in cross-domain detection. GPT-2 (Accuracy 90%, F1: 90% on tweets) a generative model, included for comparison did not perform any better than BERT and XLNet. These results suggest that the current models are not adaptable for reliable propaganda detection in different domains. This study therefore lays out a framework and baseline for future studies in this area by highlighting the limitation in a real-world cross-domain setup.

Keywords – Propaganda Detection, Cross-domain Generalization, Deep Learning, Bidirectional LSTM, Transformers (BERT, XLNet, GPT-2)

1 Introduction

Propaganda can be defined as a technique used to manipulate the public opinion by deliberately spreading misleading information. These techniques are such as cherry picking, emotive writing, logical fallacies, stereotyping, derailing from truth, selective criticism, bandwagon etc. With the advances in digital communication, it has become much easier for propagandists to set their narratives over a variety of platforms (Jowett and O'Donnell, 2018). Propaganda can be pushed in different domains with different language styles like formal news articles, informal social media posts or the old way of direct speech. This makes propaganda an interesting topic to research specifically in computational linguistics, psychology and Natural Language Processing (NLP) (Da San Martino *et al.*, 2019).

Propaganda is a very complex and sophisticated concept which involves many fine details that are overlooked intentionally or unintentionally. Prior studies have been conducted on propaganda which study one aspect of the entire thing like (Rashkin *et al.*, 2017) where they

took decorated fake news as propaganda in the study. These aspects can be studied in an isolated way, but they do not give a generalized solution to the problem. Fascinating results were seen in studies conducted by (Da San Martino *et al.*, 2019; Saleh *et al.*, 2019) on very huge news articles and social media statements but in these studies models were trained in the same domain with similar cases. While it is a stride towards understanding and tackling propaganda, fine details and psychological target on the audience of propaganda needs to be tackled by much smarter and stronger approaches and in multiple domains at the same time.

The research question in this paper is to see whether deep learning models which performed well and produced amazing results in single domain for propaganda detection are performing similarly and producing similar results in different domains or not. Deep learning models specifically transformer-based models have been chosen over the traditional machine learning models because they work well with natural language. The idea is to train the models in one domain and test them in another domain and evaluate their performance.

Further the paper contributes by revealing how the model's performance degrades on change of domain. It also puts forward possible reasons which are responsible for the degradation of model performance. The study leaves a process for evaluation and training scripts for benchmarking for researchers aspiring to work in cross-domain propaganda detection. Finally, it encourages and contributes to the United Nations Sustainable Development Goal 16 (Peace, Justice and Strong Institutions) by supporting good governance and preventing people from being manipulated by propagandists.

The paper discusses techniques and models used in propaganda detection in section 2. Section 3 describes the binary labeled propaganda data and methodology used for training and deep learning models and studying their performance. Section 4 presents the design specifications followed by section 5 which is the implementation of methods for binary classification task for propaganda detection. Section 6 reports the results on evaluation and comparison of the selected models. Finally, section 7 concludes the study and gives limitations and future work.

2 Related Work

Due to the exploding increase in propaganda on digital platforms there is an increasing interest in propaganda detection in the field of computer science. Earlier the focus for training models for propaganda detection was only on a single domain but in recent studies a generalized approach is starting to appear. Domains like news, tweets and political speeches are the kind of data suitable for learning which is studied by researchers. In this review, relevant work will be analysed, and the common points, limitations and new ideas will be presented in binary classification of propaganda detection in multiple domains. A brief comparison of the review is given in the table 1.

The study done by (Wang *et al.*, 2020) describes a cross-domain learning system trained on annotated data from speeches, tweets, and news articles. The learning was conducted on models across individual domains. The author's approach is a fusion of traditional feature-based classifiers, such as SVMs with LIWC and n-gram features, and deep learning models, such as LSTMs trained with ranking-based objectives. The authors noticed performance drop in machine-learning models for real-world applications and listed the limitations. Therefore,

this cross-domain problem motivates us to design and implement a transformer based binary classifier.

The research of (Malik, Imran and Mamdouh, 2023) on the contrary refers to highly specified hybrid feature engineering approaches. These approaches are for strong binary classifications within a particular domain. They consider a broader domain of syntactic and semantic features like character trigrams, TF-IDF, and fine-tuned BERT. The F1 scoring on the "Qprop" news dataset is extremely high. However, as this research has not been extended beyond the single domain, therefore its external validity for the results remains to be narrow. This research maintained accuracy but partially compromised generalisation, thereby highlighting the problem challenge on which this research embarked.

Shallow feature-based models provide contrasting comparisons to deep learning approaches. On the other hand, even in obtaining state-of-the-art performance on a large propaganda corpus, reliance on hand-engineered features and domain-specific tuning limited the model's flexibility. This was because the authors only employed logistic regression combined with standard preprocessing steps. This included lowercasing, stop-word removal, and vectorisation via CountVectorizer and TF-IDF (Oliinyk *et al.*, 2020) for the proposed model. A rather complicated one is the hierarchical graph-based integration network (H-GIN) proposed by (Ahmad *et al.*, 2025). The model builds bi-layered graphs encoding sequence, syntactic, and semantic information and reports more than 80 percent accuracy. The model gives results on several datasets such as ProText and PTC. Such powerful graph models however are computationally intensive and are likely brittle when applied to short informal texts such as tweets.

These limitations underscore the need for transformer-based models that provide a more adaptive solution. A solution between overly restrictive feature engineering and overly complicated graph structures. (Gupta *et al.*, 2019) provide a good example where they ensemble the merits of BERT with logistic regression and CNNs for propaganda detection at the sentence level. Their ensemble achieves state-of-the-art performance on SemEval benchmark but is still limited to news articles in a single domain. In a similar but simplified strategy, (Yoosuf and Yang, 2019) fine-tune a BERT model for binary token classification. Their system performs impressively well in detecting propaganda at the word and phrase levels, thus attesting to the ability of BERT to learn fine-grained signals in text data. However, as with (Gupta *et al.*, 2019), they restrict their evaluation to a single domain, not even attempting to explore generalization capabilities.

The generalisation domain is particularly prominent in the social media domain. In social media content the propaganda is commonly framed in short, unclear, and noisy texts with minimal noise for parsing. This is addressed in the TWEETSPIN dataset by (Vijayaraghavan and Vosoughi, 2022). It is the weakly annotated Twitter corpus for fine-grained propaganda detection. Competitive performance was achieved even in low-resource conditions. Their multi-view transformer model MV-PROP integrates contextual, semantic, and relational views. The suggestion of multi-view input and weak supervision renders reproducibility complicated in this model. Otherwise, (Barfar, 2022) has a game-theoretic model with linguistic features and LightGBM for result interpretation. Through Shapley values it attains more explainable outcome representations than complicated ones. The framework is clear

enough however it will not be tackling generalisation issues or an insistent need for representations across different platforms.

Models with high-quality training data but a narrow domain reveal other flaws. (Da San Martino *et al.*, 2019) manually annotate a news corpus with 18 propaganda techniques, training fine-grained and sentence-level detection models. The annotation is very detailed, but the system itself is strictly bound to the news domain without regard for other domains. (Sprenkamp, Jones and Zavolokina, 2023) tested prompt-engineering approaches with GPT-3 and GPT-4 on classifying propaganda techniques under zero-shot settings with competitive performance. On the downside, their models possess a tendency to output sometimes unstable or contradictory responses and thus lack the degree of reliability needed for binary classification in high stakes uses.

Overall, these works reveal a common pattern which is strong models tend to be on a single content domain. Those that seek cross-domain learning rely on architectures that are hard to scale or generalize. Even though fine-tuned transformer models have been successful within a domain, their methodical application to various data sources for binary classification has not been carried out. There is no evidence to believe that the current systems can demonstrate consistent performance in the middle of the lexical and structural diversity that typifies speech transcripts, tweets, and traditional news.

The literature therefore shows an evident gap that some models can either specialize in in-domain binary classification or do fancy cross-domain learning. And there is no thorough or systematic exploration of transformer-based binary propaganda detection across a set of domains. This study seeks to fill this gap with the creation of a transformer model trained on sentence and article level binary labelled data for news, tweets, and speeches. Also compare the performance with other transformer models and non-transformer models. It avoids dependence on rational spans or multiclass labels, thereby prioritizing simplicity and generalization. The anticipated contributions include one single binary propaganda classifier with shared representations across text domains. A measure of domain-transfer performance by conducting leave-one-domain-out experiments. The reproducibility of the benchmark architecture for future extensions. This line of research is in turn both theoretically significant and has real-world practical implications in the construction of such flexible and scalable systems for propaganda detection.

Table 1: Comparison of the literature review

Study	Domain(s)	Model / Technique	Strengths	Weaknesses
(Wang et al., 2020)	News, tweets, speeches	SVMs, LSTMs with ranking loss	Combines classical and deep learning methods	Performance drop across domains
(Malik et al., 2023)	News (Qprop)	Hybrid: TF-IDF + fine-tuned BERT	High F1 score, detailed feature set	Single domain only, overfitted
(Oliinyk et al., 2020)	News	Logistic Regression with TF-IDF, CountVectorizer	Simple and interpretable	Shallow model, poor flexibility
(Ahmad et al., 2025)	News, PTC, ProText	H-GIN (Hierarchical Graph-Based Integration Network)	High accuracy, captures syntactic and semantic cues	Heavy computation, brittle on short/noisy text
(Gupta et al., 2019)	News (SemEval)	Ensemble: BERT + CNN + Logistic Regression	Strong sentence-level performance	Restricted to one domain
(Yoosuf & Yang, 2019)	News	Fine-tuned BERT for token classification	High phrase-level detection quality	No domain transfer tested
(Vijayaraghavan & Vosoughi, 2022)	Twitter (TWEETSPIN)	MV-PROP: Multi-view transformer	Works well with weak supervision, captures different views	Hard to reproduce, limited transparency
(Barfar, 2022)	News	LightGBM + Shapley values (interpretable model)	Transparent, explainable outputs	Not focused on cross domain, weak on deep cues
(Da San Martino et al., 2019)	News (Propaganda Techniques Corpus)	Fine-grained classifier on 18 techniques	Very detailed annotation, sentence-level detection	News-only dataset, no transfer tested
(Sprenkamp et al., 2023)	Multiple (general)	GPT-3 / GPT-4 Zero-shot prompting	Flexible, zero-shot capable	Inconsistent outputs, limited binary reliability

3 Research Methodology

In this section, there is description of the dataset is followed by the description of various transformer and non transformer based models on the various combinations of datasets. These models will be trained on one domain and tested on another domain. The methodology follows four step process which is Data Collection, Preprocessing, Training and Evaluation.

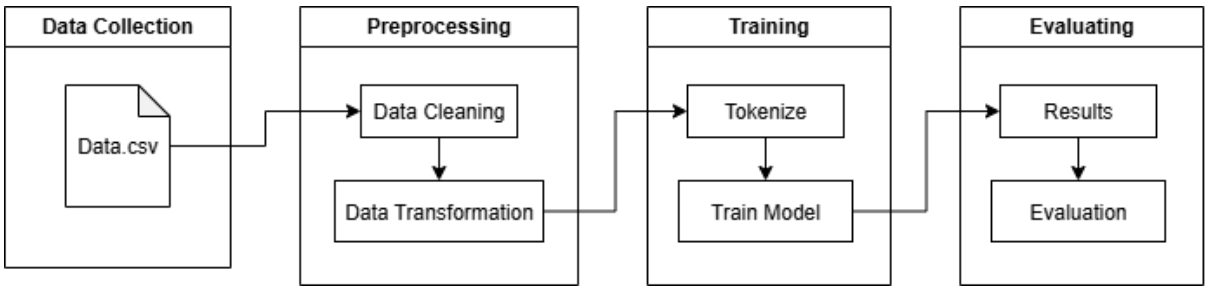

Figure 1: Research Methodology

3.1 Datasets

To achieve the goal of evaluating the deep learning models on a cross-domain setup for propaganda detection, five datasets have been used in this study from three different domains (News Articles, News Sentences, Speech Articles, Speech Sentences, Tweets). These datasets have been collected from the study of (Wang *et al.*, 2020). The raw data is available in .tsv format with extra unwanted columns which were dropped, and the new dataset was created in the .csv format. Tokenization is done a word level for GloVe lookup for LSTM and HuggingFace tokenizers for other models.

News articles and sentences: These datasets are collected from a data science competition. These are news articles and news sentences sourced from (Da San Martino *et al.*, 2019). These articles are longer texts along with binary label telling if it is propaganda or not. Similarly, the sentences are shorter text with binary labels. The articles and sentences are not sourced from the same news to maintain a cross domain learning. Both the news articles and news sentences are perfectly balanced. Average token length in table 2 shows the average number of words in the news texts in both the datasets.

Table 2: News datasets

Dataset	Total Entries	Binary Label 1	Binary Label 0	Average Token Length
News Articles	7798	3899	3899	981
News Sentences	7876	3938	3938	29

Speech articles and sentences: These datasets are collected from the speeches given by political leaders from elections of United States of America. Speeches from leaders from World War 2 are also taken. These articles are longer texts along with binary labels telling if it is propaganda or not. Similarly, the sentences are shorter text with binary labels. The

articles and sentences are not sourced from the same speech to maintain a cross domain learning. Both the speech articles and speech sentences are perfectly balanced. Average token length in table 3 shows the average number of words in the speech texts in both the datasets.

Table 3: Speech Datasets

Dataset	Total Entries	Binary Label 1	Binary Label 0	Average Token Length
Speech Articles	288	144	144	4603
Speech Sentences	24934	12467	12467	23

Tweets: This dataset is from twitter taken from Stanford University portal. It contains the propaganda tweets from Russia towards the election of United States of America. These tweets are shorter text with binary labels and are perfectly balanced. Average token length in table 4 shows the average number of words in the tweets texts in the datasets.

Table 4: Tweets Datasets

Dataset	Total Entries	Binary Label 1	Binary Label 0	Average Token Length
Tweets Dataset	17926	8963	8963	18

3.2 LSTM with GloVe Embeddings

Long Short-Term Memory (LSTM) is a model used in deep learning which is based on neural network architecture. It is used especially for sequential data like text, speech or tweets. LSTM overcame some of the major issues that were faced by some of the traditional neural networks like Recurrent Neural Networks (RNNs). One of these issues is vanishing gradients where the effects of first input diminish over time. This is important to notice because our dataset contains longer texts and therefore context needs to be maintained throughout all inputs. Another problem with the traditional models is exploding gradients which means weight are updated dramatically big on backpropagation. This makes the model unstable for tasks like propaganda detection. LSTM solves these problems by having a memory unit for saving information for longer inputs. This makes LSTM suitable for tasks like text generation, speech recognition etc. and therefore it is a suitable baseline model for the binary classification of propagandistic texts.

LSTM architecture has unique parts called the gates. These gates control the flow of information. Figure 2 shows the architecture where input gate control what new information to be added. Output gate controls what information to be passed to next step. Finally, forget gate controls what information to be dropped.

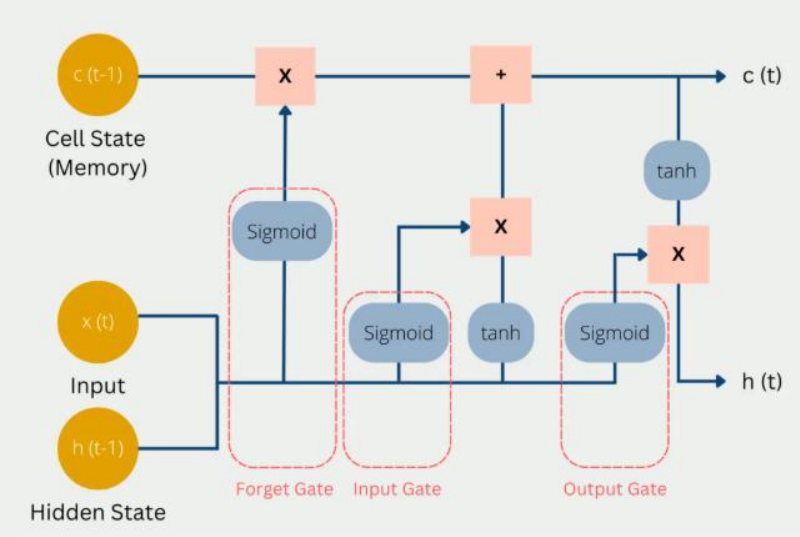

Figure 2: LSTM Architecture [1]

Without the use of transformers, the LSTM model benefits from the pretrained GloVe embeddings (Pennington, Socher and Manning, 2014) to efficiently capture the contexts of the texts which is better compared to traditional vectorizations. Evaluation strategies for LSTM Architecture for binary classification of propaganda are Accuracy, F1 Score.

Accuracy score is the most used metric for classification tasks. It gives a reliable score since the dataset is perfectly balanced. The ratio of total prediction to the true labels is given. The equation of accuracy is:

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

F1 Score is also mostly used for classification and especially binary classification. On the other hand, it combines precision and recall into one single reading. Since the model needs to properly classify both propaganda and normal text F1 Score is suitable. The equation for F1 Score is:

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN}$$

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

3.3 Transformer Models

Transformers are a neural networks architecture which was given by (Vaswani et al., 2017). It is capable of handling large sequential text data like news article or tweets. Unlike traditional neural networks the sequences are processed in parallel which is done by the self-attention mechanism. This is perfectly suitable for the classification task of given dataset as it captures long dependencies, understands the context better and scales efficiently for larger

[1] Image source: https://databasecamp.de/en/ml/recurrent-neural-network

input. It is nowadays most widely used architecture for natural language processing (NLP) and used in models like BERT, XLNet and GPT-2 etc. Figure 3 shows the architecture of transformers.

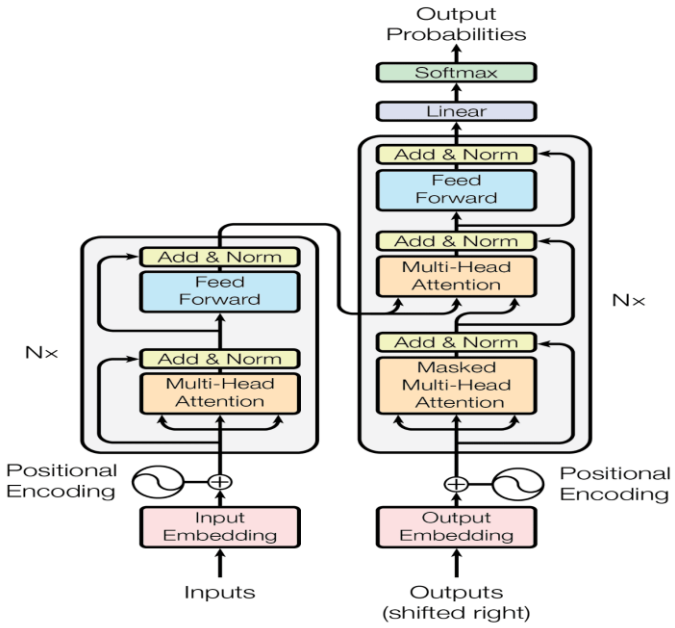

Figure 3: Transformer Architecture

Bidirectional Encoder Representations from Transformers (BERT) reads text from front and back to capture an in-depth context. BERT can be easily fine tuned to use for binary classification as required for our objective of propaganda detection. Especially the BERT is better at classifying the sentence level data. BERT is a pretrained model which makes it more powerful. This makes it is top suitable for propaganda detection hence it is considered as one of the models for evaluation.

XLNet (Yang et al., 2019) is another model considered for this study. Although BERT might be a suitable transformer-based model for binary classification of propaganda detection, the XLNet is better at modelling dependencies between words. It combines the strengths of BERT and autoregressive models. It outperforms BERT in 20 parameters and hence would be a good candidate for comparison.

GPT-2 model is generally not used for classification. It is autoregressive in nature which means it predicts the next word in a text. It is included in this study to compare ad see other models are performing better or not. With some fine tuning it can be leveraged to do binary classification.

Evaluation of these models needs metrics those are suitable with the nature of data which is subtle and nuanced.

Table 5: Metric for transformers

Metric	Why it is suitable
Accuracy	For overall performance as dataset is balanced. Can be used for bench marking.
F1 Score	Treats both the classes equally even across domains.

4 Design and Solution

The study uses an experimental design to find out how deep learning models perform for binary cross domain propaganda detection. Training model on one domain and testing it on another domain will give us the challenges faced in generalizing the models and find out which models is more robust than others.

4.1 biLSTM with GloVe

LSTM is used in bidirectional setting at true which means inputs will be processed from both directions front and back for better capture of the context. GloVe embeddings of 100 dimensions give an added benefit in the context capturing unlike earlier methods. The design for biLSTM model with GloVe for the binary classification of propaganda texts is given in figure 4. As seen in the figure the training phase is input with one domain data and on the inference phase the domain is changed.

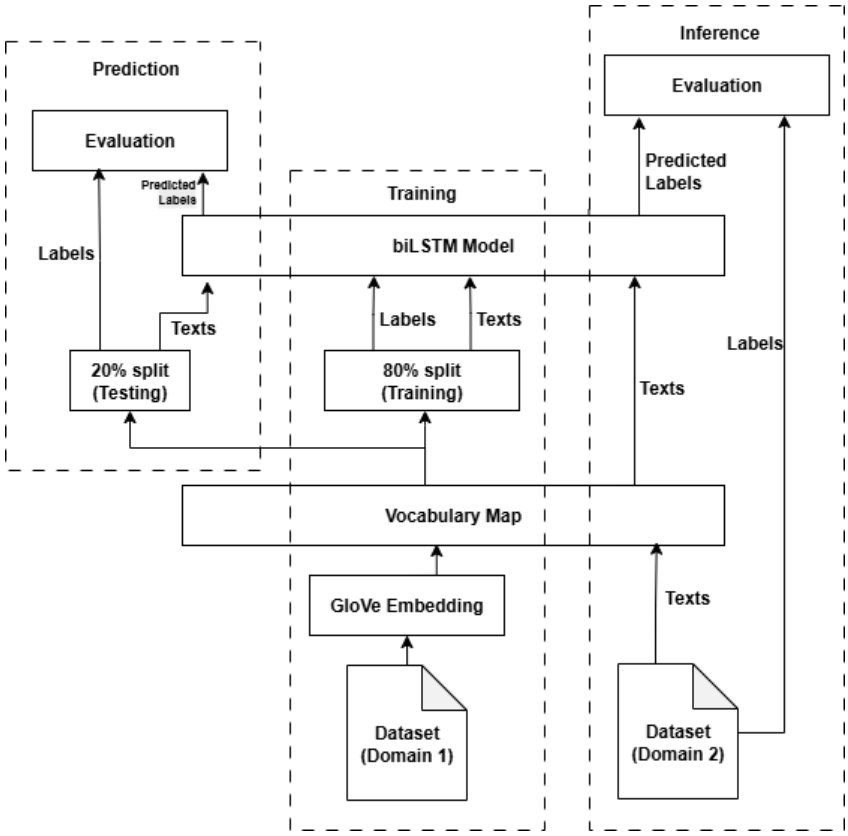

Figure 4: biLSTM with GloVe embedding model design.

The data is converted into GloVe vectors and stored in a map and split into 80% training part and 20% testing part and then passed to the two layered bidirectional LSTM. The hidden size is 128 with the batch size at 32. The training is running for 5 epoch, and loss is computed at each epoch. After the completion of training the testing is done with validation set of same domain followed by the testing on inference by another domain. The evaluation phase gets accuracy score and F1 Score.

4.2 Transformer models

The design for Transformer models is similar in all cases of binary classification of propaganda texts is given in figure 5. As seen in the figure the training phase is input with one domain data and on the inference phase the domain is changed. The transformers have their own tokenizer or encoder specific to the model. The models are pretrained on large corpora and for larger inputs chunking mechanism is adopted to avoid ambiguity.

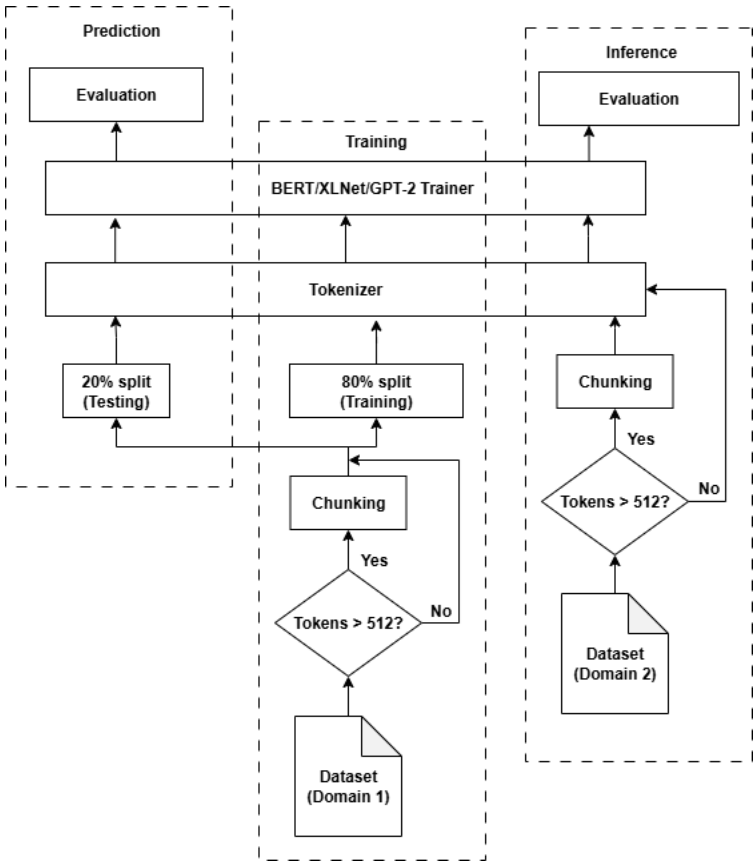

Figure 5: Transformer model design

The model BERT is a pretrained model developed by Google. It has 12 layers, 768 hidden layers and 110 million parameters. It uses BERT Tokenizer for tokenizing the input. The key training parameters to train the model for binary classification of propaganda are set using huggingface trainer API. This includes 3 epochs, a learning rate of 2e-5, batch size of 16 for both training and evaluation sets. A token limit is set at 512 tokens to make sure the ambiguity is reduced and make sure that learning is within the device limits. The accuracy and f1 score are then taken out.

The model XLNet is an autoregressive model that outperforms BERT model. It uses its own XLNet Tokenizer for tokenization of input. The key training parameters to train the model for binary classification of propaganda are set using huggingface trainer API. This includes 3 epochs, a learning rate of 2e-5, batch size of 8 for both training and evaluation sets. A token limit is set at 512 tokens to make sure the ambiguity is reduced and make sure that learning is within the device limits. The accuracy and f1 score are then taken out.

The model GPT-2 is an generative AI model that is not directly useful in the study but good as a comparison. It uses a GPT2 Tokenizer for tokenizing the inputs. It has 117 million parameters at the key training parameters to train the model for binary classification of propaganda are set using huggingface trainer API. This includes 3 epochs, a learning rate of 2e-5, batch size of 8 for both training and evaluation sets. A token limit is set at 512 tokens to make sure the ambiguity is reduced and make sure that learning is within the device limits. The accuracy and f1 score are then taken out.

5 Evaluation

The aim of the experiment was to see how well the different deep learning models perform in a cross-domain setup. Each model was evaluated in both in-domain setup and cross-domain setup. Below experiments present the accuracy and f1 score for all models. The evaluation gives a matrix showing all combinations of domains.

5.1 Experiment 1

This experiment involves the LSTM model and all domains one by one in the matrix represented in table 6.

Table 6: Evaluation results of LSTM model

Test → Train ↓	News Articles	News Sentences	Speech Articles	Speech Sentences	Tweets
News Articles	Acc: 86%, F1: 88%	Acc: 60%, F1: 72%	Acc: 65%, F1: 75%	Acc: 52%, F1: 67%	Acc: 61%, F1: 69%
News Sentences	Acc: 56%, F1: 59%	Acc: 67%, F1: 69%	Acc: 59%, F1: 64%	Acc: 48%, F1: 52%	Acc: 55%, F1: 58%
Speech Articles	Acc: 60%, F1: 48%	Acc: 55%, F1: 58%	Acc: 91%, F1: 92%	Acc: 51%, F1: 62%	Acc: 46%, F1: 54%
Speech Sentences	Acc: 50%, F1: 7%	Acc: 48%, F1: 34%	Acc: 96%, F1: 96%	Acc: 85%, F1: 84%	Acc: 52%, F1: 49%
Tweets	Acc: 64%, F1: 61%	Acc: 56%, F1: 56%	Acc: 54%, F1: 56%	Acc: 50%, F1: 49%	Acc: 81%, F1: 82%

5.2 Experiment 2

This experiment involves the BERT model and all domains one by one in the matrix represented in table 7.

Table 7: Evaluation results of BERT model

Test → Train ↓	News Articles	News Sentences	Speech Articles	Speech Sentences	Tweets
News Articles	Acc: 88%, F1: 90%	Acc: 68%, F1: 68%	Acc: 59%, F1: 73%	Acc: 56%, F1: 63%	Acc: 70%, F1: 67%
News Sentences	Acc: 63%, F1: 60%	Acc: 75%, F1: 73%	Acc: 55%, F1: 68%	Acc: 53%, F1: 43%	Acc: 51%, F1: 24%
Speech Articles	Acc: 71%, F1: 58%	Acc: 46%, F1: 33%	Acc: 93%, F1: 98%	Acc: 65%, F1: 53%	Acc: 65%, F1: 62%
Speech Sentences	Acc: 60%, F1: 10%	Acc: 57%, F1: 37%	Acc: 59%, F1: 10%	Acc: 90%, F1: 91%	Acc: 64%, F1: 53%
Tweets	Acc: 73%, F1: 73%	Acc: 56%, F1: 44%	Acc: 61%, F1: 72%	Acc: 57%, F1: 45%	Acc: 90%, F1: 90%

5.3 Experiment 3

This experiment involves the XLNet model and all domains one by one in the matrix represented in table 8.

Table 8: Evaluation results of XLNet model

Test → Train ↓	News Articles	News Sentences	Speech Articles	Speech Sentences	Tweets
News Articles	Acc: 97%, F1: 98%	Acc: 69%, F1: 73%	Acc: 59%, F1: 73%	Acc: 53%, F1: 65%	Acc: 67%, F1: 67%
News Sentences	Acc: 68%, F1: 56%	Acc: 70%, F1: 76%	Acc: 50%, F1: 53%	Acc: 60%, F1: 58%	Acc: 58%, F1: 56%
Speech Articles	Acc: 73%, F1: 71%	Acc: 55%, F1: 25%	Acc: 93%, F1: 90%	Acc: 55%, F1: 55%	Acc: 48%, F1: 5%
Speech Sentences	Acc: 62%, F1: 13%	Acc: 62%, F1: 46%	Acc: 63%, F1: 15%	Acc: 95%, F1: 91%	Acc: 64%, F1: 58%
Tweets	Acc: 63%, F1: 53%	Acc: 62%, F1: 42%	Acc: 40%, F1: 20%	Acc: 55%, F1: 21%	Acc: 89%, F1: 87%

5.4 Experiment 4

This experiment involves the GPT-2 model and all domains one by one in the matrix represented in table 9.

Table 9: Evaluation results of GPT-2 model

Test → Train ↓	News Articles	News Sentences	Speech Articles	Speech Sentences	Tweets
News Articles	Acc: 74%, F1: 78%	Acc: 56%, F1: 69%	Acc: 50%, F1: 66%	Acc: 59%, F1: 61%	Acc: 56%, F1: 60%
News Sentences	Acc: 57%, F1: 65%	Acc: 67%, F1: 68%	Acc: 53%, F1: 68%	Acc: 51%, F1: 62%	Acc: 52%, F1: 65%
Speech Articles	Acc: 47%, F1: 60%	Acc: 55%, F1: 27%	Acc: 75%, F1: 72%	Acc: 51%, F1: 10%	Acc: 46%, F1: 8%
Speech Sentences	Acc: 59%, F1: 64%	Acc: 60%, F1: 61%	Acc: 55%, F1: 61%	Acc: 70%, F1: 72%	Acc: 64%, F1: 66%
Tweets	Acc: 61%, F1: 59%	Acc: 54%, F1: 15%	Acc: 63%, F1: 69%	Acc: 55%, F1: 16%	Acc: 90%, F1: 90%

5.5 Discussion

The results from tables 6-9 show that the in-domain performance is significantly better than the cross-domain performance. XLNet tops the list in in-domain performance giving the highest accuracy of 97% for news articles. BERT is very close after XLNet giving strong performance with several accuracies over 90%. LSTM and GPT-2 on the other hand are lagging but still acceptable results (see figure 6). Speech sentences and tweets show high in in-domain results with f1 close to 90%. This shows models can adapt well on these datasets in-domain and it confirms the in-domain learning dominates.

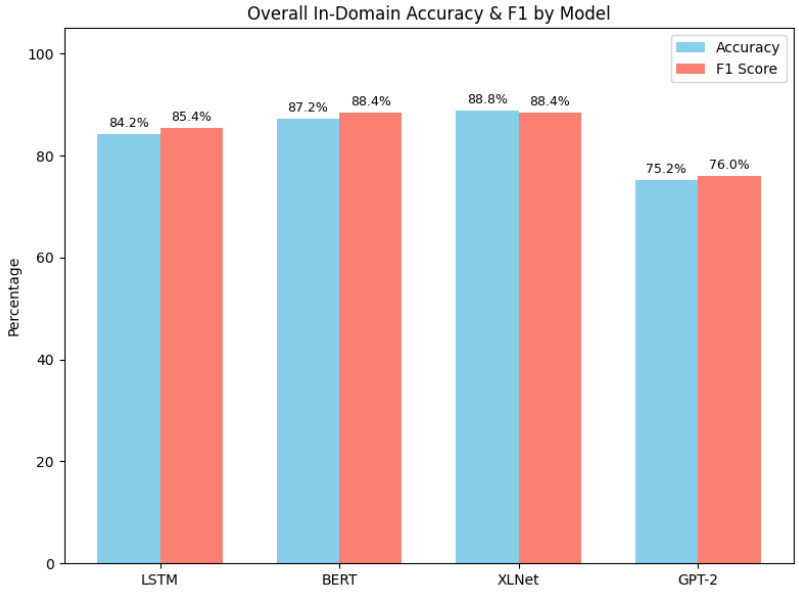

Figure 6: Overall In-domain Performance

By seeing the cross-domain results from tables 6-9, news sentences to speech sentences and speech sentences to tweets are the one which gave relatively better generalization results. This is likely due to the lengths and language style. This makes the models adaptable to these similar language structure datasets. Figure 7 shows the overall cross-domain performance.

Tweets to other domains and speech articles to tweets show the highest drop in performance where f1 drops below 20%. This shows that short lengths and inconsistent and informal language makes the model struggle to improve scores when model is trained on lengthy and formal texts.

Overall, formal and long texts are understood well in in-domain setup. Short and informal texts are better at generalizing although style and length mismatch affect negatively on cross-domain performance.

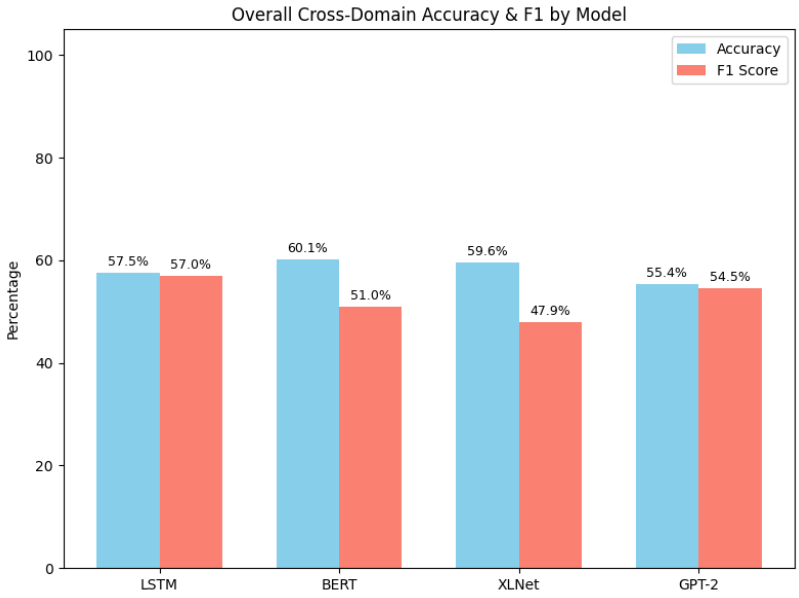

Figure 7: Overall Cross-domain Performance

6 Conclusion and Future Work

The aim of this research was to evaluate whether deep learning models can perform well for propaganda detection in multiple domains. Our objective was to determine whether models trained in one domain can maintain performance when applied to others. The experiments were done using five datasets in three domains, biLSTM model with GloVe embeddings was tested and three transformer models: BERT, XLNet, and GPT-2.

The results reveal a performance gap between in-domain and cross-domain evaluations. While all models perform strongly on the domain they are trained in, their performance deteriorates when trained in other domains. BERT and XLNet showed better cross-domain performance than the biLSTM or GPT-2. None of the models maintained consistently high accuracy or F1 scores in all domains. This demonstrates the challenge of cross-domain generalization in propaganda detection, even for advanced deep learning models.

For future work, exploring large input models like Longformer or Bigbird may yield better generalization. Additionally creating datasets with larger texts with full context could further improve model robustness and trustworthiness. Finally, models which can ignore language style and learn context would be a great step. This line of research is critical in building scalable systems for propaganda detection that align with democratic values and information integrity.

References

Ahmad, P.N. *et al.* (2025) 'Hierarchical graph-based integration network for propaganda detection in textual news articles on social media', *Scientific Reports*, 15(1), p. 1827.
Link: https://doi.org/10.1038/s41598-024-74126-9.

Barfar, A. (2022) 'A linguistic/game-theoretic approach to detection/explanation of propaganda', *Expert Systems with Applications*, 189, p. 116069.
Link: https://doi.org/10.1016/j.eswa.2021.116069.

Da San Martino, G. *et al.* (2019) 'Fine-Grained Analysis of Propaganda in News Article', in. Association for Computational Linguistics.
Link: https://doi.org/10.18653/v1/D19-1565.

Gupta, P. *et al.* (2019) 'Neural Architectures for Fine-Grained Propaganda Detection in News'. arXiv.
Link: https://doi.org/10.48550/arXiv.1909.06162.

Jowett, G.S. and O'Donnell, V. (2018) *Propaganda & Persuasion*. SAGE Publications.

Malik, M.S.I., Imran, T. and Mamdouh, J.M. (2023) 'How to detect propaganda from social media? Exploitation of semantic and fine-tuned language models', *PeerJ Computer Science*, 9, p. e1248.
Link: https://doi.org/10.7717/peerj-cs.1248.

Oliinyk, V. *et al.* (2020) 'Propaganda Detection in Text Data Based on NLP and Machine Learning', in. *MoMLeT+DS*.
Link: https://www.semanticscholar.org/paper/Propaganda-Detection-in-Text-Data-Based-on-NLP-and-Oliinyk-Vysotska/ce88a3c6b8416cefdebbf9e429ff1dee64650f00 (Accessed: 31 July 2025).

Pennington, J., Socher, R. and Manning, C. (2014) 'GloVe: Global Vectors for Word Representation', in A. Moschitti, B. Pang, and W. Daelemans (eds) *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. *EMNLP 2014*, Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543.
Link: https://doi.org/10.3115/v1/D14-1162.

Rashkin, H. *et al.* (2017) 'Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking', in M. Palmer, R. Hwa, and S. Riedel (eds) *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. *EMNLP 2017*, Copenhagen, Denmark: Association for Computational Linguistics, pp. 2931–2937.
Link: https://doi.org/10.18653/v1/D17-1317.

Saleh, A. *et al.* (2019) 'Team QCRI-MIT at SemEval-2019 Task 4: Propaganda Analysis Meets Hyperpartisan News Detection'. arXiv.
Link:https://doi.org/10.48550/arXiv.1904.03513.

Sprenkamp, K., Jones, D.G. and Zavolokina, L. (2023) 'Large Language Models for Propaganda Detection'. arXiv.
Link: https://doi.org/10.48550/arXiv.2310.06422.

Vaswani, A. *et al.* (2017) 'Attention is All you Need', in *Advances in Neural Information Processing Systems*. Curran Associates, Inc.
Link:
https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (Accessed: 5 August 2025).

Vijayaraghavan, P. and Vosoughi, S. (2022) 'TWEETSPIN: Fine-grained Propaganda Detection in Social Media Using Multi-View Representations', in M. Carpuat, M.-C. de Marneffe, and I.V. Meza Ruiz (eds) *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. *NAACL-HLT 2022*, Seattle, United States: Association for Computational Linguistics, pp. 3433–3448.
Link: https://doi.org/10.18653/v1/2022.naacl-main.251.

Wang, L. *et al.* (2020) 'Cross-Domain Learning for Classifying Propaganda in Online Contents'. arXiv.
Link: https://doi.org/10.48550/arXiv.2011.06844.

Yang, Z. *et al.* (2019) 'XLNet: Generalized Autoregressive Pretraining for Language Understanding', in *Advances in Neural Information Processing Systems*. Curran Associates, Inc.
Link: https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html (Accessed: 6 August 2025).

Yoosuf, S. and Yang, Y. (2019) 'Fine-Grained Propaganda Detection with Fine-Tuned BERT', in A. Feldman et al. (eds) *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*. *NLP4IF 2019*, Hong Kong, China: Association for Computational Linguistics, pp. 87–91.
Link: https://doi.org/10.18653/v1/D19-5011.