Classifying English and Arabic news articles using Prompt-Based Techniques

Tanmay Vakare, Alekhya Hari, Sriraj Vuppala, Karan Meda, Muktan Patel, Devang Vamja, Dhruvi Sonani, Priyansh Suthar.

Abstract:

The objective of this study was to enhance the performance of language models in identifying political violence by fine-tuning BERT, RoBERTa using prompting techniques. Two types of prompting were employed: tuning-free prompting and fixed prompt language model tuning. The results showed that fixed prompt language model tuning was the most effective approach, achieving an average F1-score of 0.784 across both models on the BBC violence classification dataset. These findings highlight the potential of prompting techniques in enhancing the ability of language models to analyze and identify political violence in news articles. The study also applied the same fine-tuning techniques on a Arabic language dataset for fake news classification and demonstrated that fixed prompt language model tuning was more effective, resulting in an average F1-score of 0.62.

CCS Concepts: Computing methodologies → Artificial intelligence, Natural language processing, Classification and Prompt-based Learning.

Keywords : Prompting, Language Models, BERT, RoBERTa.

Code: https://github.com/Tanmay06/promptingPolitical

1. Introduction

Recent advances in natural language processing (NLP) and machine learning have enabled the development of language models that can automatically identify and analyze patterns in large volumes of text data. In the field of political science, these models hold promise for enhancing our ability to monitor and analyze instances of political violence.

Several studies have explored the use of NLP techniques for identifying political violence in news articles. For instance, researchers have used topic modeling, sentiment analysis, and network analysis to identify patterns of violence in news reports. However, these methods often rely on predefined categories or keywords, which can limit their effectiveness in detecting subtle and complex patterns of violence.

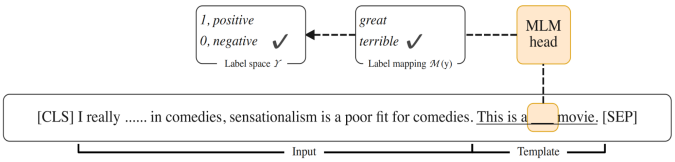
Prompting is a technique that has emerged as a promising approach for enhancing the performance of language models in various NLP tasks, including identifying instances of political violence. Prompting involves providing a short piece of text as input to the language model, which serves as a prompt for generating relevant responses.

In this project, we leverage prompting techniques to improve the performance of BERT, RoBERTa in classifying news articles. Specifically, we explore two types of prompting: tuning-free prompting and fixed prompt language model tuning.

Tuning-free prompting involves no tuning. Fixed prompt language model tuning, on the other hand, involves predefining a set of prompts that are specifically tailored to the task of identifying political violence for English dataset and identifying fake news for Arabic dataset.

Through our experiments, we seek to determine which approach is most effective in enhancing the performance of the language models on this task. Our findings demonstrate the value of leveraging prompting techniques in enhancing the ability of language models to identify and analyze instances of political violence.

The remainder of this paper is organized as follows: In Section 2, we detail our proposed approach for enhancing language models' ability to classify news articles using prompting techniques. In Section 3, we describe the experiments we conducted, and in Section 4, we present the observations we made based on our findings. Finally, in Section 5, we offer our conclusions based on our work.



2. System Design

2.1 Prompt Engineering

Prompt Engineering is the process of designing and fine-tuning prompts or templates to improve the quality of language models' results. Main idea behind the concept is to guide the model to generate more relevant, coherent, and accurate responses. This process involves creating a set of structured prompts or input templates that the language model can use to generate responses that meet specific criteria.

The key benefit of using prompt engineering is that even when you train a language model on a comparatively small dataset, it can improve the model's output generation efficiency. This is because it guides the model towards generating more coherent and accurate responses, even when it has limited context to work with. We have included 43 simple and complex prompts (both Prefix and Cloze Task combined) to implement the module.

- Prefix Task Examples:
  - **$x** news is **$y**
  - In news **$x** the incident type is **$y**

    - In the news **$x**, the incident reported was **$y**
- Cloze Task Examples:
    - **$x** This is a **$y** news
    - In the news story, **$x**, an incident of **$y** occurred.
    - The incident of **$y** was reported in the news story **$x**.

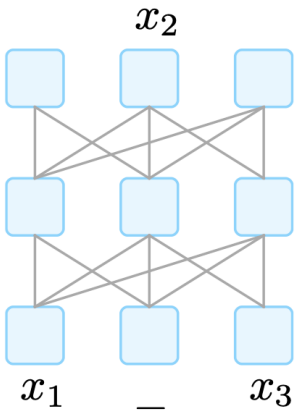
2.2 Pretrained Language Model

Since we are experimenting with both cloze and prefix prompts for classification, bidirectional autoregressive models like BERT and RoBERTa are good pretrained models for our task. The Masked Language Modeling task can be adopted with prompts for classification.

- **Bert**

  BERT, stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers, is a highly effective and widely used transformer based language model.

  BERT is designed to **pretrain deep bidirectional representations** from unlabeled text by jointly conditioning on both left and right context in all layers. Further it can be **finetune on variety of NLP tasks** (NER, Classification, MNLI)



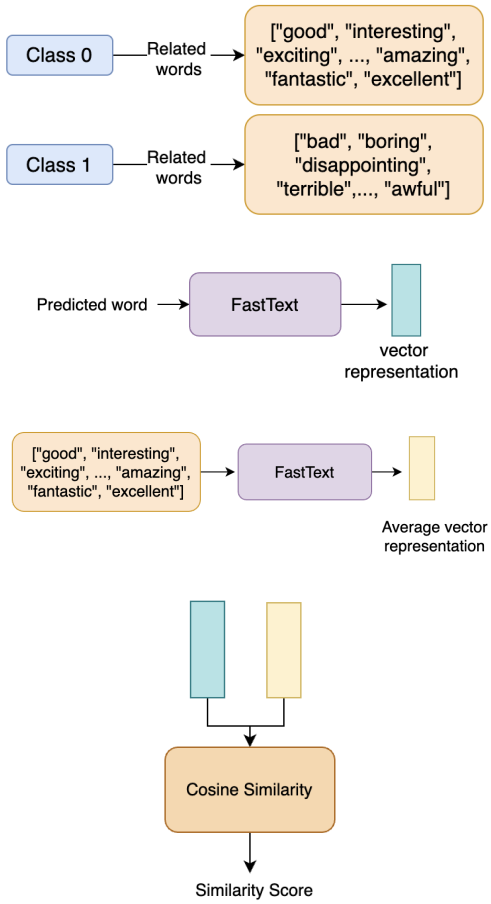
- **RoBERTa**

  RoBERTa(Robustly Optimized BERT Approach) is a variant of BERT that incorporates several improvements to further enhance its performance. The key improvements in ROBERTA include:

  1. Training on More Data
  2. Dynamic Masking
  3. Larger Model Size

2.3 Answer Engineering

Answer Engineering is the process of creating a set of predefined responses or answers to a given set of questions or prompts. Which involves identifying relevant keywords and phrases that represent the classes or categories of responses that the Language model should provide.



2.4 Dataset Description

1. **English Dataset:** The **BBC_News**[1] **dataset** is a collection of news articles that have been annotated with binary class values, specifically 0 and 1, indicating the presence or absence of a political conflict within the news article. The dataset is a curated collection of diverse news articles that contain information regarding various aspects of politics, such as elections, legislative decisions, protests, and armed conflicts, among others. The dataset is designed to be used for training and evaluating natural language processing models that aim to detect political conflicts in news articles.

News Article	Class	Train/Valid/Test
tv future in the hands of viewers with home theatre ...when they want.	0	1588/315/322

2. **Arabic Dataset:** AFND[2] (**A**rabic **F**ake **N**ews **D**ataset) is a group of public Arabic news articles collected from public Arabic news websites. Which is composed of 606912 news articles collected from 134 different public news websites, where each one is stored as a separate sub-directory in the "Dataset" named directory. Each of them has a json file which contains a JSON object containing title, text and publication date of the article.

News Article]	Class	Train/Valid/Test
"ترأس عبد القادر اعمارة، وزير التجهيز و النقل و اللوجيستك و الماء الجمعة الجلسة الافتتاحية للندوة الدولية المدن المغربية و عدد من الفاعلين في المجتمع المدني"	credible	1363/292/293

For both dataset we identify the set of keywords that represent each class label for eg. in english dataset ["good", "excellent", "amazing", ..., "fantastic"] are the keywords corresponding to label 0. These keywords will help us to finetune the model and direct it towards correct token prediction. This process is described in detail in the experiment section.

3. Experimental setup

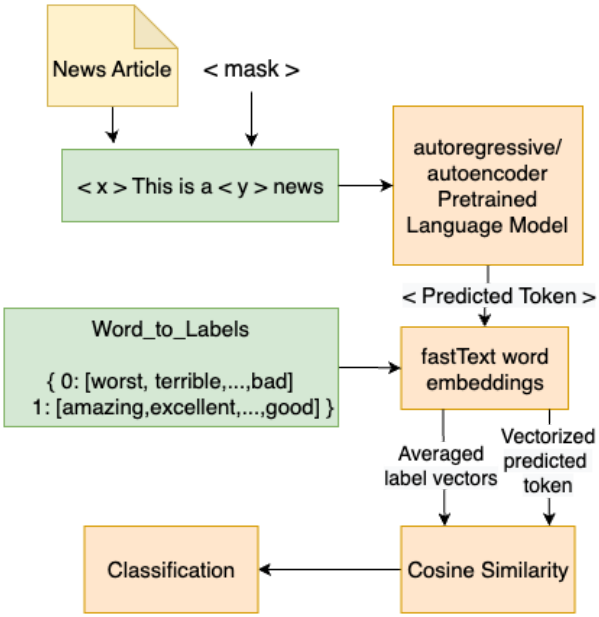
3.1 Prompting Tasks

a. Tuning-free prompting :

Tuning-free prompting is an efficient technique that allows pre-trained language models to generate answers without modifying their parameters. It relies on prompts for generating responses, and in-context learning combines prompt augmentation with tuning-free prompting. This approach prevents parameter updates and catastrophic forgetting, making it useful in zero-shot settings. However, it requires significant engineering effort to achieve high accuracy, and generating multiple responses during testing can be slow and impractical, limiting scalability.

b. Fixed Prompt LM tuning

Fixed-LM prompt tuning is a method that updates prompt parameters while keeping the pre-trained language model unchanged. It is effective for few-shot scenarios, retains the model's knowledge, and often outperforms tuning-free prompting in terms of accuracy. However, it cannot be used in zero-shot scenarios, has limited representation power for large datasets, requires prompt engineering, and produces prompts that are not easily interpretable or manipulable by humans. In conclusion, fixed-LM prompt tuning has advantages but also limitations that should be considered in its application.



3.2 Proposed Approach

1. Build input sentence by appending prompt template at the end of News article
2. Set mask token in Prompt template so that our MLM model predicts the possible tokens for mask
3. Using average vector of each label and vector representation of predicted token and find cosine similarity between them
4. Assign the predicted token to respective class

4. Results

According to the analysis BertForMaskedLM (bert-base-uncased) with Fixed Prompt LM Tuning reports better results among all experiments below.

Model	Task	Dataset	Accuracy	f1
RobertaForMaskedLM (roberta-base)	Tuning Free Prompting	BBCNews	74	0.1333
BertForMaskedLM (bert-base-uncased)	Tuning Free Prompting	BBCNews	18	0.305
RobertaForMaskedLM (roberta-base)	Fixed Prompt LM Tuning	BBCNews	88	0.7499

BertForMaskedLM(bert-base-uncased)	Fixed Prompt LM Tuning	BBCNews	92	0.8181
RobertaForMaskedLM(xml-roberta-base)	Tuning Free Prompting	Arabic Fake News Dataset (AFND)	48	0.4901
BertForMaskedLM(aubmindlab/bert-base-arabertv2)	Tuning Free Prompting	Arabic Fake News Dataset (AFND)	46	0.5972
RobertaForMaskedLM(xml-roberta-base)	Fixed Prompt LM Tuning	Arabic Fake News Dataset (AFND)	58	0.6121
BertForMaskedLM(aubmindlab/bert-base-arabertv2)	Fixed Prompt LM Tuning	Arabic Fake News Dataset (AFND)	62	0.6415

Best LM output result after tuning, where yellow is the prompt and green is the predicted token:

{'score': 0.05666811764240265, 'token': 4206, 'token_str': ' excellent', 'sequence': 'defence cordon was slowly disintegrating. england prop matt stevens ran in at full steam to suck in a few more tacklers. unfortunately he ran into o connell who hit him hard - very hard - and then wrestled the ball away for a crucial turnover. that spoke volumes about ireland s back-foot display with defensive coach mike ford taking a bow at the end. to win a game like that showed that ireland have moved forward. it may be tries that win games but it is defence that wins championships. This is a excellent news'}

{'score': 0.9893759489059448, 'token': 101208, 'token_str': 'حقيقي', 'sequence': '."إقناع الناس بالمشاركة أو ردع المتجاهلين للانتخابات" وأشار براهمي إلى أن وزارة العدل حاولت سابقاً تمرير قانون مخالف لحقوق الإنسان متعلق بنزع الجنسية، لولا تدخّل الرئيس الذي أوقف المشروع، لافتاً إلى أن الحريات عبارة "عن ممارسة وليست مجرد آلية فقط". وكلّف الرئيس عبد المجيد تبون الحكومة توفير الظروف المناسبة لإجراء الانتخابات المقبلة وتأمينها، والسماح للجزائريين باختيار ممثليهم في البرلمان بكل حرية، فيما تعهّد الجيش بتوفير كل الظروف الآمنة لإجراء الانتخابات، والتصدي لكل من يسعى إلى عرقلتها. هذه أخبار حقيقي'}

5. Conclusion

In this Project, we identified that across both English and Arabic languages, the news classification task demonstrated significantly improved performance when using Fixed Prompt LM Tuning compared to Tuning Free Prompting. The results showed an average increase of 28.5% accuracy and an average increase of 0.324 on f1 score. However, with improved prompts structure models can generate better results as well as improved answer engineering can reduce false token classification.

Reference

1. BBC News Classification - https://www.kaggle.com/c/learn-ai-bbc/overview
2. Arabic Fake News Dataset (AFND) - https://data.mendeley.com/datasets/67mhx6hhzd/1
3. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - https://arxiv.org/abs/1810.04805
4. RoBERTa: A Robustly Optimized BERT Pretraining Approach - https://arxiv.org/abs/1907.11692
5. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing - https://arxiv.org/pdf/2107.13586v1.pdf
6. ConfliBERT: A Pre-trained Language Model for Political Conflict and Violence - https://aclanthology.org/2022.naacl-main.400.pdfx