Classifying English and Arabic news articles using Prompt-Based Techniques

Team 12

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

Outline

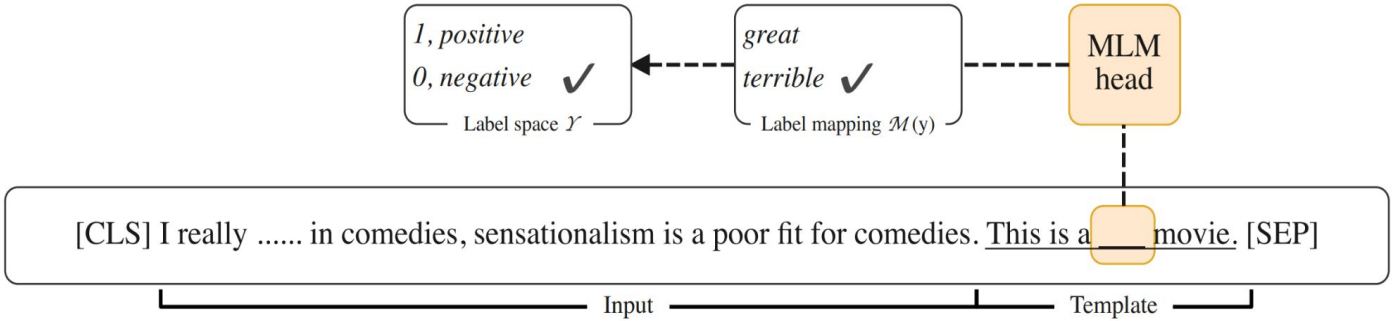
- ➔ Introduction
- ➔ Why prompting?
- ➔ Prompt Engineering
- ➔ Pre-Trained Language Model
- ➔ Answer Engineering
- ➔ Dataset
- ➔ Experimental Setup
- ➔ Proposed Approach
- ➔ Results
- ➔ Examples
- ➔ Conclusion

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 and classify large volumes of text data.
- However, these methods often rely on predefined categories or keywords, which can limit their effectiveness in detecting subtle and complex patterns of violence.
- Hence we utilize prompting for classification of English and Arabic news articles using Masked Language Models (MLMs) Bert and Roberta.

Why prompting ?

- 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.



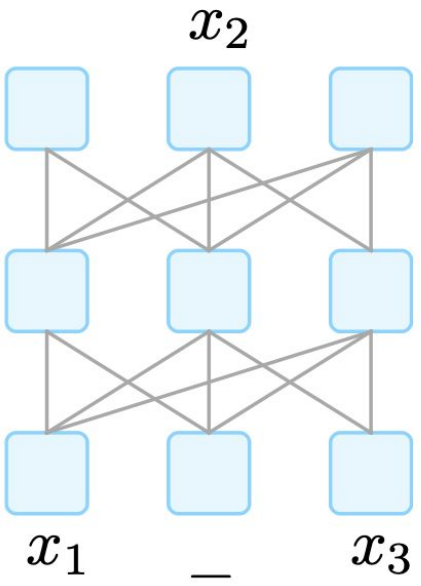
- With limited data and few-shot learning prompting can give comparatively good results.

Prompt Engineering

- Prompt engineering is important for correctly directing PLM.
- Engineered 43 simple and complex prompts.
- Prefix Task
  - **$x** news is **$y**
  - In news **$x** the incident type is **$y**
  - In the news story **$x**, the incident reported was **$y**
- Cloze Task
  - **$x** This is a **$y** news
  - In the news story, **$x**, an incident of **$y** occurred.
  - The incident of **$y** was reported in the news story **$x**.

Pre-Trained Language Model

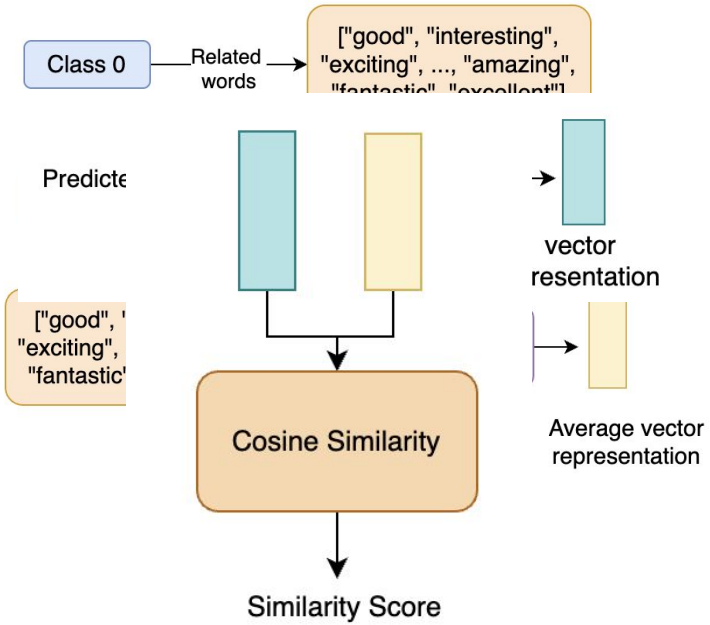
- Used auto-regressive Masked Language Model for downstream classification task.
- BERT [1]
  - BERT, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers, is a highly effective and widely used transformer based language model
- RoBERTa [2]
  - RoBERTa has the same architecture as BERT, but uses a byte-level BPE as a tokenizer and uses a different pretraining scheme.



[1] https://arxiv.org/abs/1810.04805
[2] https://arxiv.org/abs/1907.11692

Answer Engineering

- Found out keywords for representing the classes
- Transformed the predicted token and the keywords to context-free word representations using fastText [3]
- Calculated the cosine similarity between the token and the average of keywords.
- Mapped the token to the class higher cosine similarity.



[3] https://arxiv.org/abs/1607.04606

Dataset

- The **BBC News Binary classification dataset [4]** 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 article.

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

- (AFND) [5] is a collection of public Arabic news articles that were collected from public Arabic news websites. Misbar, which is a public Arabic news fact check platform, is used to classify the articles into credible, not credible, and undecided.

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

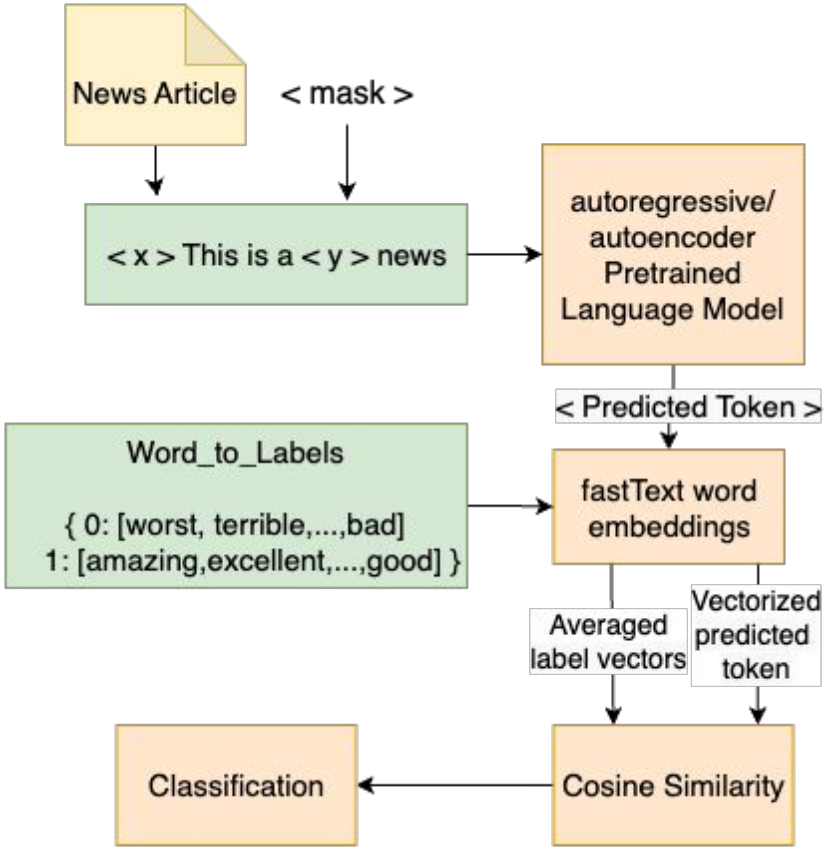
[4] https://www.kaggle.com/c/learn-ai-bbc/overview
[5] https://data.mendeley.com/datasets/67mhx6hhzd/1

Experimental Setup

- Tuning Free Prompting
  - A technique that enables the generation of answers directly from pre-trained language models (LMs) without the need to modify their parameters.
  - Advantages:
    - Zero-shot
  - Disadvantages:
    - Heavy prompt engineering
- Fixed Prompt LM Tuning
  - Fixed-prompt LM tuning tunes the parameters of the LM, but additionally uses prompts with fixed parameters to specify the model behavior.
  - Advantages:
    - Few-shot
    - Better task specific performance
  - Disadvantages:
    - Better task specific performance

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



Results

Model	Task	Dataset	Accuracy	f1
RobertaForMaskedLM (roberta-base)	Tuning Free Prompting	BBCNews	74	0.1333
BertForMaskedLM (bert-base-uncased)	Tuning Free Prompting	BBCNews	18	0.3050
RobertaForMaskedLM (roberta-base)	Fixed Prompt LM Tuning	BBCNews	88	0.7499
BertForMaskedLM (bert-base-uncased)	Fixed Prompt LM Tuning	BBCNews	92	0.8181
RobertaForMaskedLM (xlm-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 (xlm-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

Example

- {'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'}

Conclusion

- Fixed Prompt LM Tuning performed better than Tuning Free Prompting on both the languages English and Arabic for news classification task with average of **+ 28.5%** accuracy and average **+ 0.324** on f1 score
- With better engineered prompts the model can give better results.
- Better mapping/answer engineering can reduce the false classification of tokens

References

[1] https://arxiv.org/abs/1810.04805

[2] https://arxiv.org/abs/1907.11692

[3] https://arxiv.org/pdf/2107.13586v1.pdf

[4] https://www.kaggle.com/c/learn-ai-bbc/overview

[5] https://data.mendeley.com/datasets/67mhx6hhzd/1

[6] https://aclanthology.org/2022.naacl-main.400.pdf

Thank You

Team 12

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