Exploring Topics of the Mid and Post Crisis Greek Reality Through Prime Ministerial Speeches

Nikolaos Giannakopoulos and **Dionysios Rigatos**
Norwegian University of Science and Technology
Athens University of Economics and Business
{nikolagi, dionysir}@stud.ntnu.no

1 Project Theme

This project aims to analyze the thematic content of speeches and statements delivered by Greek Prime Ministers over the past twelve years (2012-2024). This has been an eventful period for Greece as it coincides with a significant economic crisis, refugee scandals, natural disasters, and the COVID-19 pandemic - all while alternating between left and right wing governments in a late-and-post-crisis Greece. We will make use of topic modeling techniques so as to hopefully identify recurring and relevant topics as well as analyze this data scientifically in order to draw conclusions regarding the state of affairs in modern day Greek politics.

2 Data

Our dataset consists of official transcripts of speeches and statements given by the office of the Greek Prime Minister during the selected period. Those transcripts can be found directly on the official website of the Greek PM(1). The data will be manually collected using our own customized improvement over an existing Greek PM speech scraper(2). Each speech/statement transcript will be considered a separate data point and we might include additional metadata such as date, prime minister, and political party. All the data will undergo the standard preprocessing needed for most natural language processing tasks. Preprocessing will include, but will not be limited to, lowercasing, tokenization, stopword removal, lemmatization as well as anything else deemed necessary. Libraries such as SpaCy-EL(3) or the gr-nlp-toolkit(4) are at our disposal for preprocessing.

3 Methodology

The following is subject to slight modification or change as the assignment progresses, as issues come up and better techniques are discovered. However the main ideas regarding how we are planning to tackle this task for latent topic identification are:

- **Latent Dirichlet Allocation** as a baseline model for our task using techniques shown in the lecture as well as adjustments required for task compatibility.
- **BERTopic**(5) as a State-of-the-Art model. While BERTopic is multilingual, we would like to experiment with swapping out the built-in sentence transformer for a Greek-specific sentence transformer such as the Greek Media SBERT(6) for improved results. Further experimentation will be decided on the go.
- **Zero-Shot Modeling** with the labels extracted using the aforementioned models (with manual editing if need be) so as to see whether we can achieve similar or better results with this technique. A candidate for this is the nli-xlm-r-greek(7) model.
- **Any other technique** that might naturally occur or replace any of the aforementioned on our discretion.

Further techniques such as use of different word embeddings and dimensionality reduction techniques (PCA, UMAP, Linear Discriminant Analysis) will be decided on a need-to basis.

4 Evaluation

Evaluation unsupervised tasks such as topic modeling is inherently difficult. Our task's evaluation criteria consists of, but is not limited to, the following techniques:

- **Coherence Score** for evaluating topic coherence, reflecting topic quality by measuring the degree of semantic similarity between high scoring words within each topic.
- **Manual analysis** for evaluation of the results in combination with the preliminary EDA.
- **Visualization** of the topics so as to help us make conclusions regarding their similarity.
- **Other quantitative measures** such as silhouette score (for BERTopic), KL-Divergence etc.

Topic modeling usually requires a lot of manual qualitative evaluation. For quantitative evaluation, machine learning or evaluation-specific libraries such as **OCTIS**(8) offer methods for performing the aforementioned in a reliable and consistent manner.

5 Alternative Project Ideas

While we are confident that this topic is interesting, we offer the examiner alternative ideas we considered during the preparation of this phase - should they deem that the main topic should be changed. Main alternatives were:

- **Neural machine translation for European languages** using the EU Parallel Translation Corpus(9).
- **Greek court tranascript summarization** using the GreekLegalSum(10) dataset.

References

[1] https://www.primeminister.gr

[2] https://github.com/kritonp/primeminister-speeches-scrapper

[3] https://spacy.io/models/el

[4] https://github.com/nlpaueb/gr-nlp-toolkit

[5] https://maartengr.github.io/BERTopic/index.html

[6] https://huggingface.co/dimitriz/st-greek-media-bert-base-uncased

[7] https://huggingface.co/lighteternal/nli-xlm-r-greek

[8] https://github.com/MIND-Lab/OCTIS

[9] https://www.kaggle.com/datasets/hgultekin/paralel-translation-corpus-in-22-languages

[10] https://huggingface.co/datasets/DominusTea/GreekLegalSum