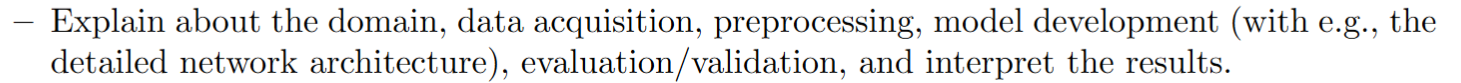
**INTRODUCTION**



Text classification is a common task to do in the field of machine learning and natural language processing (NLP). This task usually consists of applying preprocessing techniques to the input dataset such as tokenization, stemming, and stop-word removal, followed by applying machine learning or deep learning algorithms such as random forests, support vector machines (SVM), or transformer-based models like BERT.

In this report, the author has attempted to build and train a classifier that’s able to categorize scientific paper abstracts into one of three predefined scientific fields using the BERT (Bidirectional encoder representations from transformers).

**BACKGROUND**

**BERT**

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model based on an “encoder-only” transformer architecture. To understand BERT, it is important to understand the transformer architecture.

Transformers

**METHODOLOGY**

**Dataset Acquisition, Cleaning, & Preprocessing**

The author acquired 300 abstract datasets for **all four scientific fields** for experimentation purposes. It was acquired by scraping journals specific to a given field using **CrossRef API** with the script being written in Python. CrossRef API allows access to metadata of scholarly articles, including abstracts.

The scraper retrieves abstract data for a given journal by making a request to CrossRef API, filtering the results based on a specified **journal name** to find abstracts in each field. It then processes the response and extracts available abstracts and saves them to a CSV file each with a label corresponding to the field of the journal (e.g., "sociology"). The scraper performs this for each scientific field until it gathered at least 300 abstracts. Here is the list of journals used for each field:

* **Astronomy**: The Astronomical Journal, Publications of the Astronomical Society.
* **Political Science**: Annual Review of Political Science, British Journal of Political Science.
* **Psychology**: Annual Review of Clinical Psychology, Psychological Science.
* **Sociology**: Sociology from British Sociological Association, Annual Review of Sociology, American Journal of Cultural Sociology.

After scraping, cleaning the datasets involved removing:

* Duplicates.
* Unicode characters.
* Empty entries such as “Not available” or “” (empty strings).
* HTML artifacts such as <jats:p>...</jats:p>.
* Whitespaces, line breaks, tabs and inconsistent formatting.
* Text artifacts such as “Abstract:” and “Conclusion:” to prevent interference with training. Since certain journals included these artifacts while others did not, the network might inadvertently learn to associate these terms detracting from its ability to focus solely on the abstract content.

This makes sure that the datasets only contain texts relevant to the content of the abstract.

One thing to note was that the author performed data acquisition before deciding on which models to use for text classification. Therefore, it included steps that are unnecessary when using BERT such as:

* Unicode cleaning, BERT typically handle Unicode characters without issues.
* Whitespaces, BERT tokenizers ignore extra spaces during tokenization.

Transformer-based models such as BERT handles preprocessing during tokenization.

**EXPERIMENTATION**

**RESULTS**

**CONCLUSIONS**