News Categorisation: Text Classification of Titles and Excerpts of News

Kai Sun Anson Lam

[redacted]

1 Introduction

In this coursework, I intend to categorise news in the selected dataset with text classification techniques of Natural Language Processing (NLP). On account of the increasing amount of data generated, while digital news have become more easily accessible, this has created the challenge of identifying and finding contents of ones' interest (Kaur and Bajaj, 2016). News categorisation, hence, has gained significant importance to assist internet users to have efficient access to news of their interest in real time (Kaur and Bajaj, 2016). In an attempt to categorise the news in the dataset, text classification of titles and corresponding excerpts of news with various NLP models would be useful.

2 Background

Since news categorisation, to a large extent, is based on text classification, further understanding of specific text classification techniques would be useful. Kowsari et al. (2019) illustrated an overview of text classification pipeline, including feature extraction, dimensionality reduction (optional), classification (i.e. learning model), and evaluation (i.e. prediction of test data and evaluation of model). Classification, obviously, forms the most crucial part of the whole pipeline. Provided with multiple news categories as classes in the dataset, Support Vector Machine (SVM), particularly multi-class SVM, is the first candidate for classification technique, ranging from One-vs-One to All-vs-One variant to be possibly implemented (Kowsari et al., 2019). Another classification approach is Neural Network (NN), including Deep Neural Network (DNN), Recurrent Neural Network (RNN), and Convolutional Neural Network (CNN) to be considered (Kowsari et al., 2019). Finally, in order to evaluate the models, a number of evaluation metrics, for instance accuracy, sensitivity, specificity, precision, and recall, can be used (Kowsari et al., 2019).

3 Proposed methodology

- **Review the identified problem:** News categorisation is a text classification problem. With respect to the dataset, a couple of NLP models are selected as main approaches to categorise the news through analysing the titles and excerpts. With Support Vector Machine (SVM) and Neural Network (NN) being the two potential candidates of approaches, the clear focus of models means that the implementation and comparison of them is of medium difficulty.

- **Conduct a literature review:** Kaur and Bajaj (2016) specifically illustrated the workflow of news categorisation, reviewed a multitude of algorithms particular for news categorisation, and analysed the pros and cons of different approaches of relevant algorithms. The workflow includes 4 main stages: news collection, news pre-processing, feature selection, and news classification (Kaur and Bajaj, 2016), which is in line with the more general text classification pipeline of Kowsari et al. (2019) mentioned in the Background.

- **Develop a plan**: Based on the pipeline of text classification of Kowsari et al. (2019) and the workflow of news categorisation of Kaur and Bajaj (2016) in the literature review, I have a more specific plan for the coursework. With the selected the dataset as detailed in 3.1, the first step is data pre-processing, such as tokenisation, stop word removal, and word stemming (Porter stemmer), if necessary. In terms of feature extraction or selection, I intend to use relative frequency, bigram, and perhaps product of these two features. Depending on the dataset, principal component

analysis might be used to reduce the number of dimensions. As for specific classification techniques, Support Vector Machine (SVM) and Neural Network (NN) are two major approaches. With regards to the evaluation method, accuracy is the main evaluation metric to be used, though others might be considered as well.

- **Explain what the contributions would be**: As I mentioned in the Introduction, the main goal of news categorisation models to be developed is to enable internet users to have access to desired news contents more easily and efficiently. In addition to being applied to the dataset, the models are aimed to categorise unseen news and be used for larger corpora in the future.

- **Anticipate project duration**: The duration of the entire project is expected to be within 8 weeks, given that news categorisation, specifically text classification of titles and excerpts in the dataset, is of medium difficulty. Details of the proposed timeline is set in 3.3.

3.1 Data

The dataset to be used is consisted of over 5,500 unique news articles and corresponding excerpts in English language, with 4,686 rows of training set and 828 rows of testing set by default. Data fields include 'Title' (string of news title), 'Excerpt' (string of short extract from news body), and 'Category' (string of category of Title and corresponding Excerpt as label). Provided with the 'Category' data field as label, including 'business', 'sports', 'politics', 'health', 'entertainment', and 'tech', annotation is not required. This publicly accessible dataset is available on Hugging Face at https://huggingface.co/datasets/okite97/news-data.

3.2 What baselines are you considering?

Two baselines are considered to be compared with the main approaches, Support Vector Machine (SVM) and Neural Network (NN). The first baseline is the majority label baseline. This baseline aims to predict the most frequent label in the training set, without taking input features into account. Another baseline is the bag of word baseline. This baseline intends to represent the frequency of each word, which appears in documents ('Title' or 'Excerpt'), and belongs to the corpus.

3.3 Proposed timeline

- Data pre-processing (news tokenisation, stop word removal, and word stemming (Porter stemmer) etc.) – 1 week
- Feature extraction (relative frequency, bigram, product of these two features, etc.) – 2 weeks
- Baselines development (majority label baseline and bag of word baseline) – 1 week
- Main models development (Support Vector Machine (SVM) and Neural Network (NN)) – 2 weeks
- Prediction and evaluation (Accuracy etc.) – 1 week
- Spare time – 1 week

4 Experimental setup and tools

Although documents in the dataset are all in English language, pre-processing on data mentioned above would ensure the input of models are of equal standard. The two classification models to be developed are Support Vector Machine (SVM) and Neural Network (NN). Regarding Support Vector Machine (SVM), 'svm' from Scikit-learn is expected to be used. Considering Neural Network (NN), tool options include TensorFlow, PyTorch, and Scikit-learn. During the model development process, Google Colaboratory is assumed to be the main platform.

References

Gurmeet Kaur and Karan Bajaj. 2016. News classification and its techniques: a review. *IOSR Journal of Computer Engineering*, 18(1):22–26.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4):150.