MSc Project Proposal – School of Computer Science, University of Birmingham

Name: Kenya Williams

Project Title: Beyond the COVID-19 Clickbait: A Comparative Analysis of Supervised and Unsupervised Models for Headline and Article Similarity

Relevance to Programme (Data Science): This project leverages data science by applying supervised and unsupervised classification algorithms to a real-world challenge: combatting COVID-19 misinformation. Using Python and data science libraries, I will build and evaluate models, ensuring responsible data management practices. Effective visualisations will illustrate the impact of each approach.

Abstract

This project investigates the effectiveness of supervised and unsupervised learning models in identifying similarities between COVID-19 news headlines and articles. By comparing these models, the study aims to enhance the reliability of news consumption and empower readers to recognise potentially sensational news sources. The results will be visualised to clearly communicate the comparative performance of each approach.

1. Introduction

News consumers require access to reliable information for informed decision-making; however, misleading headlines, particularly prevalent in news about the COVID-19 pandemic,[1] make it difficult to discern trustworthy information. Focusing on the similarity between news headlines and articles can play a crucial role in helping users navigate the information-rich yet attention-grabbing online news environment [2]. This project explores which supervised and unsupervised learning models are most effective at identifying headlines and content that discuss the same underlying events.

2. Project Requirements

I. Data Acquisition
Collect a comprehensive dataset of COVID-19 related headlines and articles from various reliable and representative news sources.
Preprocess the data to clean, normalise, and structure it for analysis.

II. Model Development
Select and configure both supervised (e.g., SVM, Random Forest) and unsupervised (e.g., K-Means) learning models, with text embedding techniques (e.g. BERT) and deep learning architectures (e.g. CNNs) for similarity analysis.
Establish a baseline for comparison and evaluate models using existing and scraped datasets, defining a fit threshold.

III. Similarity Analysis and Comparative Evaluation
Conduct a comparative analysis using metrics like accuracy, precision, recall, and F1-score to evaluate model performance (supervised vs. unsupervised, text embedding techniques, deep learning architectures).
Design visualisations (e.g. confusion matrices, bar charts) to effectively communicate the comparative performance of models.

IV. Ethical Considerations
Address and document ethical considerations related to data collection, usage, and analysis to ensure compliance with ethical standards.

3. Literature Review

Existing research in computational media explores various methods for headline similarity analysis. Some models focus on topic detection by timeline (Allan, 2012) but this approach is often too broad to distinguish between sub-topics discussed in small timeframes, such COVID-19 reporting. Others (Wubben et al., 2009) are more capable but fail to verify whether their algorithms correspond with human labelling and annotations. The few attempts at unsupervised Natural Language Understanding (Labal et al., 2021) struggle to compete with the best supervised learning methods. A key gap lies in the lack of balanced comparisons between these models, and most studies solely focus on headlines, neglecting the richer context of full articles. Additionally, existing datasets may not capture the nuances of regional news coverage.

This project addresses these limitations by conducting a comprehensive comparison of supervised and unsupervised models, incorporating both headlines and article content. Utilising a diverse UK-based COVID-19 news dataset, it aims to identify the most effective models and features for similarity detection. This will contribute to the development of more reliable news similarity analysis tools, empowering users to navigate the media landscape with greater discernment. Through highlighting dissimilar headlines discussing the same underlying event, this project can also indirectly uncover potentially sensational news sources.

4. Methodology

This project adopts a systematic approach to ascertain the best models for detecting similarity between COVID-19 news headlines and content. The process begins with data collection: a labelled training dataset and a separate scraping of UK-based COVID-19 news sources will provide data for training and manual testing. Data pre-processing, including text cleaning, normalisation, and structuring will prepare the data for analysis. Next, supervised models (e.g., Support Vector Machines, Random Forest) and unsupervised models (e.g., K-Means clustering) will be implemented to assess similarity. Methods for accurately evaluating models will include metrics such as accuracy, precision, recall, and F1-score. Feature extraction techniques and potentially BERT may be used to enhance similarity assessment. Finally, visualisations created with Matplotlib and Seaborn will communicate the comparative strengths and weaknesses of each model. This comprehensive approach identifies the most effective techniques for COVID-19 news similarity analysis.

5. Project Management

This project is expected to be completed within 2 months, otherwise separated into 8 weekly milestones. Proposed timeline:

- **Week 1:** Pre-process labelled training data, perform initial analysis (EDA).
    - **Milestone:** Ready training data
- **Week 2:** Identify and scrape UK news sources, clean and analyse collected data.
    - **Milestone:** Cleaned and analysed testing data
- **Week 3:** Select and develop supervised learning models for similarity analysis.

    - **Milestone:** Functional supervised model prototypes
- **Week 4:** Select and develop unsupervised learning models for similarity analysis.
    - **Milestone:** Functional unsupervised model prototypes
- **Week 5:** Test model performance on a small test set from labelled dataset, refine models.
- **Week 6:** Define evaluation metrics, evaluate all models' performance.
    - **Milestone:** Model evaluation complete
- **Week 7:** Test models with manual test set (COVID-19), create data visualisations for model comparison.
- **Week 8:** Finalise report, prepare presentation summarising findings.
    - **Milestone:** Project completion

Resources:

- Data Source: HLGD Headline Grouping Dataset [3]
- Python libraries (pandas, sci-kit learn, NLTK)
- Visualisation libraries (Matplotlib, Seaborn)
- Text embedding libraries (TensorFlow, PyTorch)

Deliverables:

- Usable code repository
- Functional model prototypes (supervised & unsupervised)
- Presentation summarising key findings
- Final report with detailed analysis

References

- [1] Statista. (n.d.). *Coronavirus fake news frequency in the UK 2020*. [online] Available at: https://www.statista.com/statistics/1112492/coronavirus-fake-news-frequency-in-the-uk/.

- Colomé, J.P. (2024). *Misleading headlines in mainstream media are more dangerous than outright fake news*. [online] EL PAÍS English. Available at: https://english.elpais.com/technology/2024-05-30/misleading-headlines-in-mainstream-media-are-more-dangerous-than-outright-fake-news.html#https://english.elpais.com/technology/2024-05-30/misleading-headlines-in-mainstream-media-are-more-dangerous-than-outright-fake-news.html# [Accessed 28 Jun. 2024].

- Allan, J. (2012). *Topic Detection and Tracking*. Springer Science & Business Media.
- Sander Wubben, Antal Van Den Bosch, Emiel Krahmer, and Erwin Marsi. 2009. Clustering and matching headlines for automatic paraphrase acquisition. In Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009), pages 122–125.

- [3] paperswithcode.com. (n.d.). *Papers with Code - News Headline Grouping as a Challenging NLU Task*. [online] Available at: https://paperswithcode.com/paper/news-headline-grouping-as-a-challenging-nlu-1 [Accessed 28 Jun. 2024].