News Popularity Prediction Project Proposal

In the news media business, news popularity is highly related to publisher’s revenue and audience engagement. Accurately predicting the popularity of newly published articles would help publishers optimize content creation and enable them to focus on topics and stories that resonate better with their audience. Therefore, the objective of this project is:

**To predict news popularity based on article’s headline and body using feature extraction approach.**

In this project proposal, we outlined a method to develop a news popularity prediction model based on the features extracted from articles headlines and body.

Multiple datasets will be used in this project. The main dataset, [Online News Popularity](https://archive.ics.uci.edu/dataset/332/online+news+popularity)1, is provided by UCI including urls to the original contents and number of shares to estimate its popularity. The dataset includes about 40,000 news articles. The headlines and body will be scraped by urls. Other datasets like [News Sentiment Dataset2](https://www.kaggle.com/datasets/hoshi7/news-sentiment-dataset) and [Title and Headline Sentiment Prediction](https://www.kaggle.com/datasets/adarshsng/title-and-headline-sentiment-prediction)3 will be used to improve the sentiment analysis feature extraction. [News Category Dataset](https://www.kaggle.com/datasets/rmisra/news-category-dataset)4 will be used to create category features based on headlines. All dataset used in this project is publicly available.

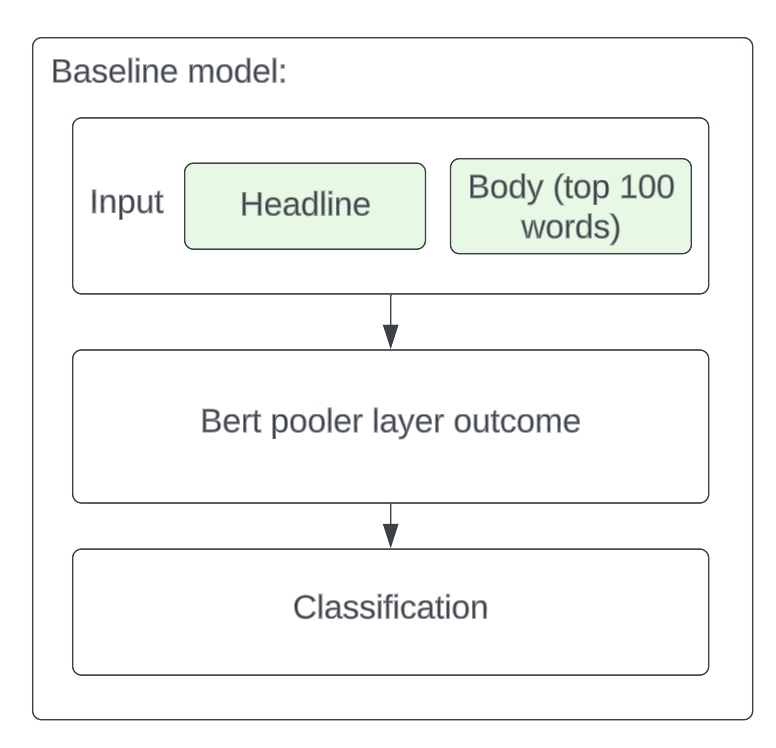
Literature review:

This topic and dataset is well studied by researchers. The original dataset includes some article metadata and article features but not the original text data. However the method to extract those features is not provided. Syed7 worked on multiple regression models based on the original dataset and achieved 68.3% as the highest accuracy with an ensemble of logistic regression, LDA and Random forest model. The top 5 important features are number of embedded links, number of images, number of videos, amount of subjectivity in title and amount of subjectivity in contents. However, those features are not available in other articles. Therefore the model cannot generalize to other contents. [Anuja](https://onlinelibrary-wiley-com.libproxy.berkeley.edu/authored-by/Arora/Anuja)8 proposed a method that adds text features from article title and body text plus image in addition to original text. The text feature drive are readability features, eg. Lix score and smog index which is popular among editor’s lenses. The issue with this combined method is that not all articles had an image attached.

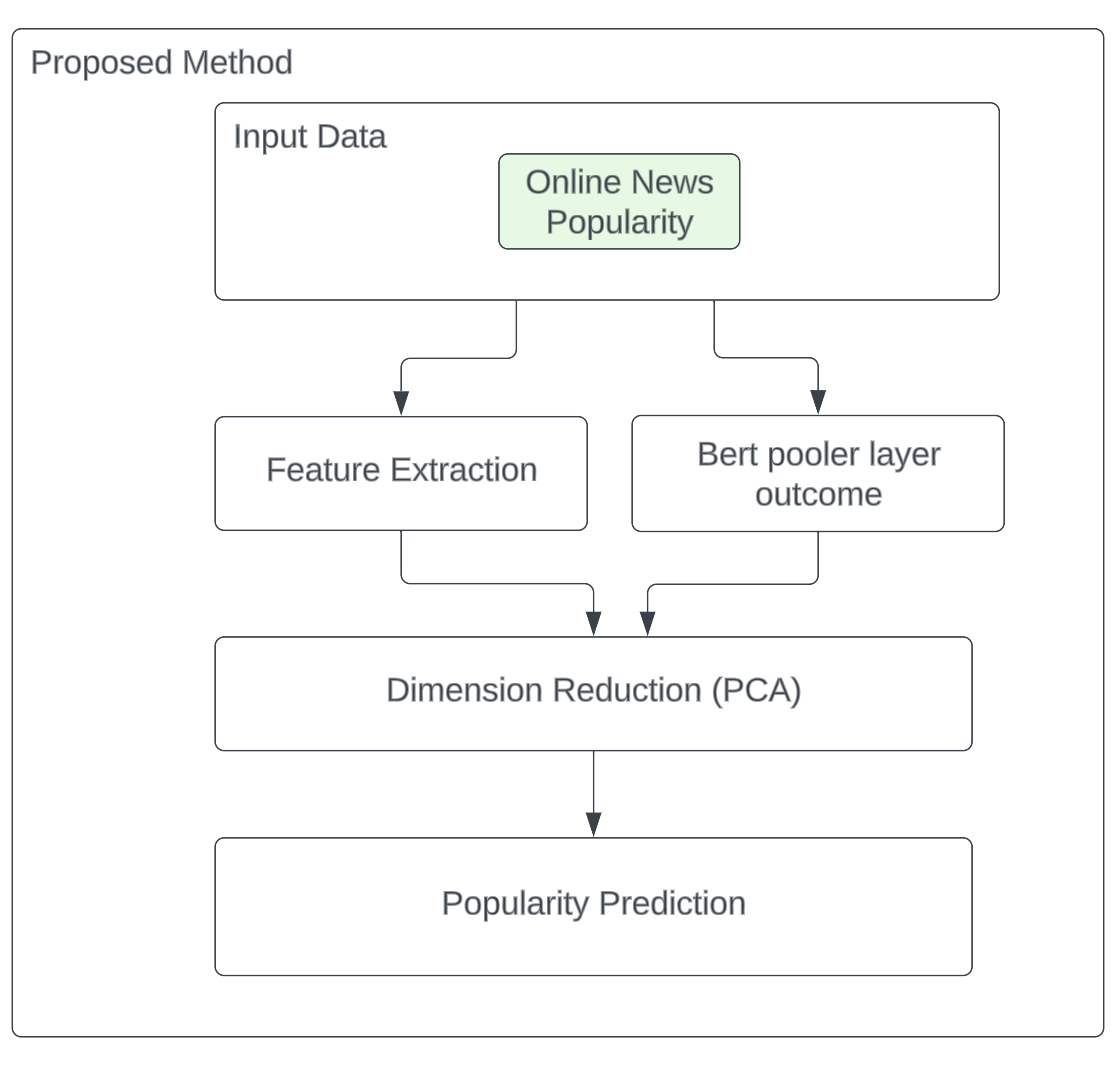
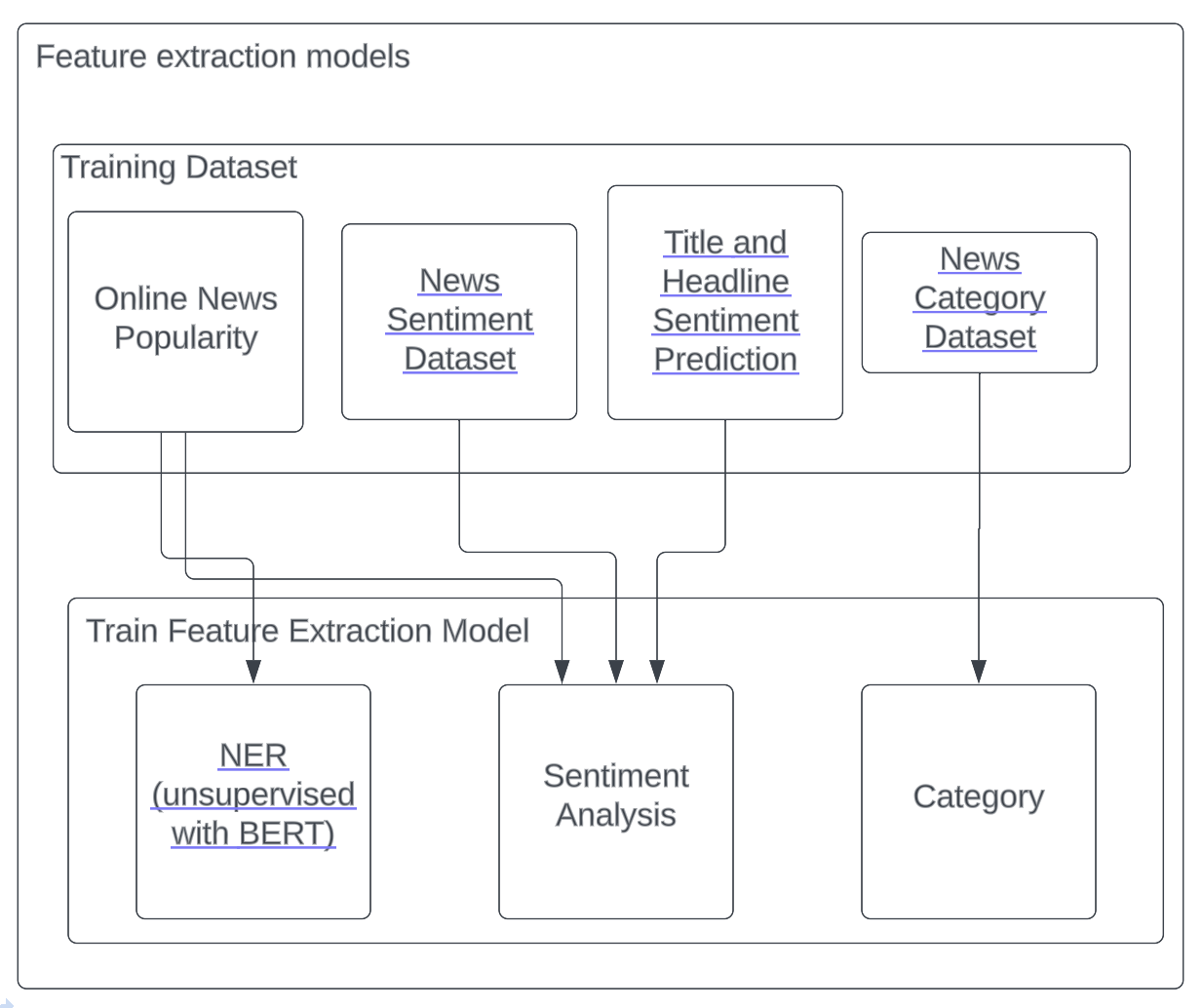
Ramsha9 proposed a framework to predict early news popularity based on text features such as sentiment, topic popularity, engagement features. The model is trained on various publisher’s articles therefore it generalizes well. Liao10 proposed a deep fusion method to combine Temporal Process (time series) and Content Features (text feature and metadata). Liao suggested the Content Features fit better to predict article popularity right after its publication (within a few hours). The Temporal process fits better with long term popularity. Most of previous research suggests that sentiment and article categories are the most important and effective features when predicting the popularity of news articles. Therefore, in my proposed method, sentiment and article categories are included. There is no study on including NER features to build a regression or classification model. The NER features are similar to many features suggested by Ramsha9 and the main difference is that it will be done in a single feature engineering step.

Therefore, in my proposed method, I will combine NER, sentiment, article categorization features, basic metadata and BERT pooler layer to create a regression model.

The baseline model would take headlines and body as input. Using BERT as an embedding method and followed by a classification model to predict if the content will be popular or not. The figure below demonstrates the baseline model design:



The proposed method will include three additional feature extraction techniques on top of the BERT pooler layer such as U[nsupervised NER using BERT](https://towardsdatascience.com/unsupervised-ner-using-bert-2d7af5f90b8a#:~:text=NER%20is%20done%20unsupervised%20without,location%20on%20CoNLL%2D2003%20corpus.)5, [Sentiment Analysis](https://towardsdatascience.com/sentiment-analysis-on-news-headlines-classic-supervised-learning-vs-deep-learning-approach-831ac698e276)6 and Category detection. Each feature extraction would be a separate model. Combining 3 extracted features and the BERT pooler layer as classification model input to complement input data. The hypothesis is by adding relevant features, the model will outperform the baseline model. The figure below demonstrates the proposed model’s feature extraction method and classification model design:



Reference:

1. <https://archive.ics.uci.edu/dataset/332/online+news+popularity>
2. <https://www.kaggle.com/datasets/hoshi7/news-sentiment-dataset>
3. <https://www.kaggle.com/datasets/adarshsng/title-and-headline-sentiment-prediction>
4. <https://www.kaggle.com/datasets/rmisra/news-category-dataset>
5. <https://towardsdatascience.com/unsupervised-ner-using-bert-2d7af5f90b8a#:~:text=NER%20is%20done%20unsupervised%20without,location%20on%20CoNLL%2D2003%20corpus>.
6. <https://towardsdatascience.com/sentiment-analysis-on-news-headlines-classic-supervised-learning-vs-deep-learning-approach-831ac698e276>
7. <https://medium.com/@syedsadiqalinaqvi/predicting-popularity-of-online-news-articles-a-data-scientists-report-fac298466e7>
8. <https://doi-org.libproxy.berkeley.edu/10.1111/exsy.13336>
9. <https://doi.org/10.1016/j.eswa.2021.116496>
10. Liao, Dongliang et al. “Popularity Prediction on Online Articles with Deep Fusion of Temporal Process and Content Features.” *Proceedings of the ... AAAI Conference on Artificial Intelligence*. Vol. 33. N.p., 2019. 200–207. Web.