SENTIMENT ANALYSIS OF JOURNALISTIC NEWS ARTICLES USING BILSTM

ALEXANDER TAN KA JIN

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Your Degree

Faculty Of Computing
Universiti Teknologi Malaysia

21/12/2024

Methodology

Methodological Framework

A research methodological framework allows a researcher to illustrate the core steps in conducting a research and helps the reader to understand the given steps the researcher took to receive their end results. This aids in replication of the research and helps identify potential criticisms of the research. This thesis will follow four main steps.

1. Research Gap Identification
2. Data Collection
3. Data Preprocessing
4. Model Construction
5. Model Performance Evaluation
6. Model deployment

Through literature review, research gaps are found and the purpose of this study is constructed. We determined that this research will encounter the following challenges in it's methodology:

1. Determining the correct dataset and ensuring that the proper features are extracted from the data prior to model training.
2. Choosing the correct model and training method. Which can greatly impact the final behaviour of the model and the results of this research. If training is supervised or semi-supervised, then the data labelling method must be considered too.
3. Choosing a dataset to analyse relationships between politics opinions and emotional sentiment.
4. Determining the proper visualizations such that relationships between data are made clear.

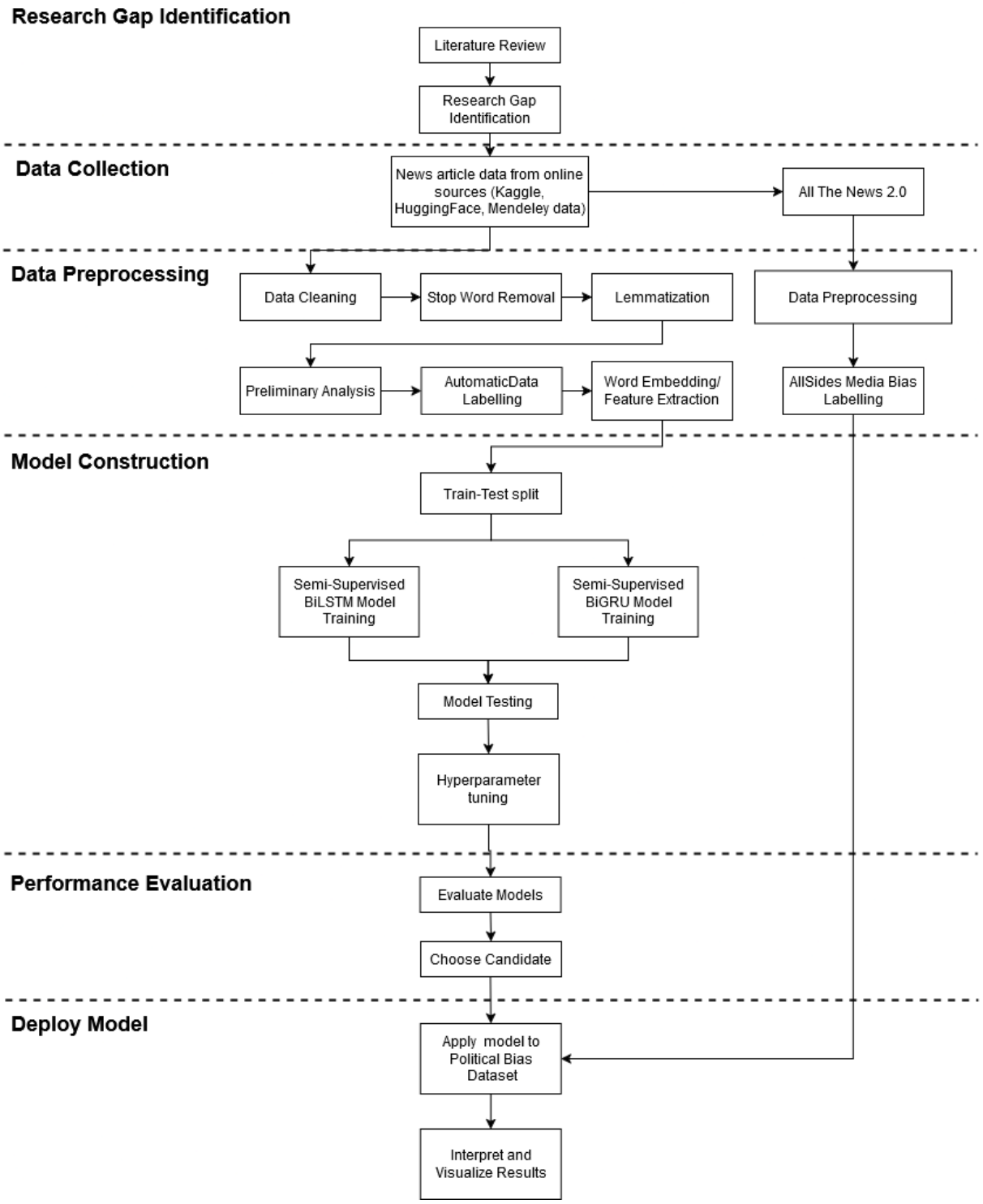

Figure 1 Illustration of the research framework

Data Collection

The subject of this study is long-form text in the form of news articles, therefore the data used has to be same. Most articles are freely available on internet and are often scraped by users on the net and published through Kaggle and Mendeley Data. One notable article is the All The News 2.0 Dataset (Thompson, n.d.) which is a large dataset containing 2.6 million articles from news of different websites including author, title, main text and publication. The size and diversity makes it perfect for use in analysis of news media.

The main obstacle to data collection is the diversity and size of the new article dataset. Preferably, the project requires a large dataset of atleast 1 million articles and sufficient diversity of data such that all kinds of news outlets are covered regardless of bias and regardless of the factuality of it's contents. Searching on sites like Kaggle and HuggingFace, some potential sources of training/testing data were identified, the most preferable candidate is most likely the Reddit Ideological Bias Dataset due to it's large size and diverse articles:

Dataset Name	Dataset Description	News source
Reuters 21578 (David D. Lewis & Li, 2004)	10,369 documents of Reuters news articles. One of the most popular text datasets for text classification and catergorization research.	Reuters
BBC Articles Dataset with Extra Features (Ferretti, n.d.)	2,127 BBC news articles with text and labels for multiple categorization and classification use.	BBC
BBC News (Preda, n.d.)	A large dataset of around 40,000 BBC news articles. Contains main text and meta-data. Self updating.	BBC
NYT Articles: 2.1M+ (Singh, n.d.)	A very large dataset of 2.1 million New York Times articles with meta-data, updated daily.	NYT
Reddit Ideological bias datset(Ravi & Vela, 2024)	A large dataset consisting of articles shared on reddit which contains a large diversity of news from various outlets including both real and fake news.	Various

Table 1 Dataset candidates

Model and training Choice

Prior to model construction, it is important to both chose the model and chose the training method that is being used for research. For this research, BiLSTM is chosen as it has been proven in the literature review to have promising results in sentiment analysis. BiGRU is similar and offers slightly lower performances but is more lightweight and would also be chosen alongside BiLSTM to compare performances. The bidirectionality of these models provide them the capability to analyse sentences in context of the sentences before and after and has been proven in Literature Review to have great results in sentiment analysis for shorter forms of text.

The type of model greatly determines the methods of data preprocessing and how much of the data is labelled. For most of sentiment analysis, supervised analysis is used. (Wankhade et al., 2022) noted that semi-supervised training may be more effective at handling ambiguous use of language while supervised training is better at handling subjectivity. Since most of the data will contain objective use of language, large amounts of subjective language is not of concern and semi-supervised training will be used.

Data Preprocessing

Data Preprocessing for text involves feature extraction so that the accuracy of the model is increased and labelling so that we can direct the model into doing sentiment analysis.

Data Cleaning

Prior to preprocessing the data, we must ensure that the training/testing data does not contain any irrelevant noise that may decrease the accuracy of the model. For this research, the model must properly interpret professional English journalism and so flaws that may impede this process are:

1. Hyperlinks (https:// or http://)
2. Stop words (a,the,in)
3. Internet Slangs (btw, aka, omg)
4. Punctuations (, . @ !)
5. Unusually capitalized words/ lowercase words (sUCh aS THIS)
6. Misspelt words (Suchhh es tis)
7. Misplaces spaces/misattached words (suchas t his)
8. Non-English words
9. Reduplicated words (...Department of Statistics Malaysia (DOSM)...)

In addition, missing values, invalid values (such as a text being too short) and satirical articles (such as The Onion and The Babylon Bee) are removed. Any institution or entity that has an acronym should be turned into acronyms as to emphasize the entity rather than it's individual word components. Words that are not the root of a word should then be lemmatized (eg. Happily becomes happy, mournful becomes mourn).

After the data is cleaned for every dataset, those that are chosen for training are concatenated into one singular training/testing dataset. Any data outside of the main text is discarded (title, date of publishing, author, etc.)

Data Labelling

For the training/testing dataset, the data is labelled as to prepare for supervised training using automatic labelling software such as VADER or TextBlob. This project will use SentiWordNet 3.0 (Baccianella et al., 2010) as it has the additional benefit of identifying between objective and subjective statements.This allows the project to properly filter out facts that are identified to be purely neutral and objective which are abundant in journalism. Of the training data, only a handful of data will be labelled

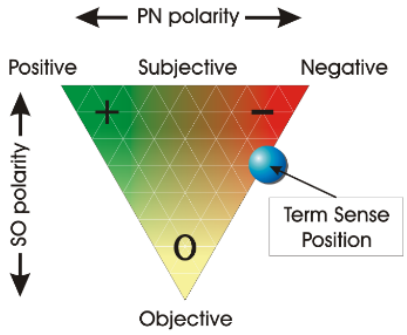

Figure 2 Sentiment classification criteria for SentiWordNet. It classifies based on how objective/subjective is it and for subjective opinions how strongly positive/negative. (Sebastiani & Esuli, 2006)

during semi-supervised training which is detailed in the Model Training and Data Split section of the methodology. All testing/validation data will be labelled.

Word Embedding

The main methods of transforming words into vectors are documented within the literature review. Between the three options, Word2Vec is preferable as to avoid potential computational constraints that may be met when running GloVe. Word2Vec is comparably lightweight compared to GloVe while still extracting features based on context.

The vectors from the result of Word Embedding will then be normalized with

min-max scaling as to ensure that data is standard and uniform while maintaining the relationships between vectors.

Model Training and Data Split

The data will be split between Training, Testing and Validation on a 80/10/10% split. Training data is used to help the model learn and encode patterns within data, the testing data is used to evaluate the performance of the model and the validation set is used to tune hyper-parameters of the model. Within semi-supervised training, only a small part of the training data is labelled while the rest is pseudolabeled. Pseudolabelling is the process of labelling unlabelled training data and then adding it back to the training data on the next session. Only 10% of the training data will labelled this way and the remaining are pseudolabelled.

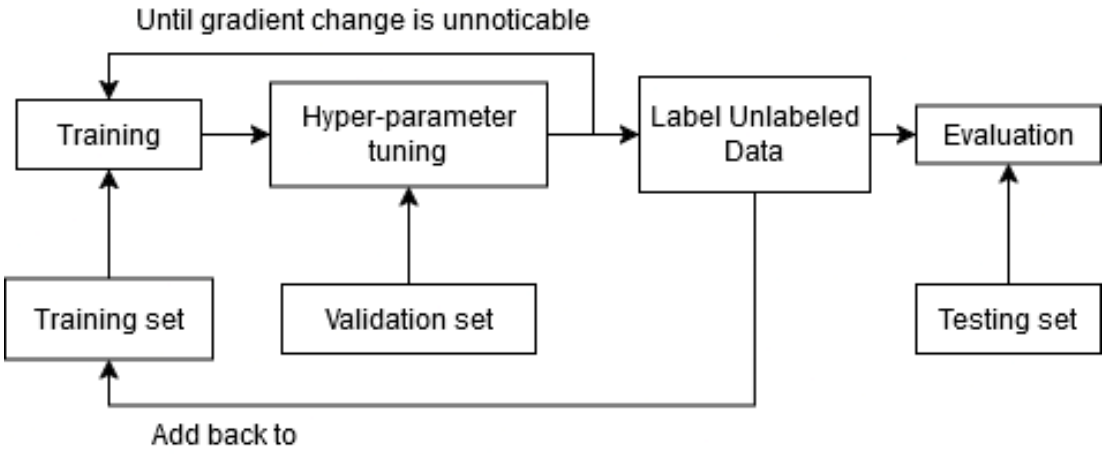

Figure 3 Semi-supervised training and evaluation process

When a model is done training, it moves to the hyper-parameter tuning section of the phase where the validation set is used to adjust variables like learning rate and optimiser momentum. After which, it loops back to training until gradient change of the loss function is too little for any noticeable improvement.

As noted in model choice, the models used for this research will be BiLSTM

and BiGRU. The two will be trained simultaneously and at evaluation, the better performing model will be chosen to be deployed.

Evaluation

The end model will be a classification model. A classification model labels data between two or more categories. Classification models are often evaluated using a confusion matrix which measures the amount of labels that were predicted correctly and incorrectly. The confusion matrix lists down the true positive (top left), false positive (top right), false negatives (bottom left) and true negatives (bottom right). Subsequent evaluation metrics like Recall, Precision, Accuracy and F1 score is derived from the initial confusion matrix. Due to the research's use of SentiWordNet, the confusion matrix used is 3x3 to evaluate positive, negative and neutral responses.

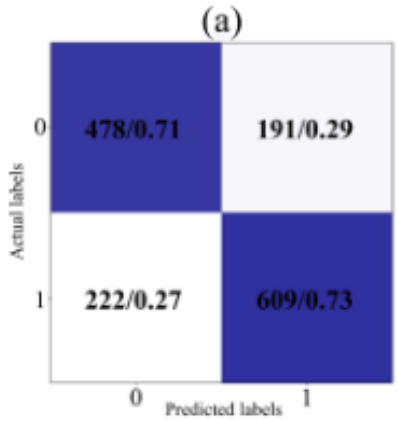

Figure 4 Example of a Confusion matrix(Mu et al., 2024)

Below are the formulae for Accuracy, Precision, Recall and F1 scores respectively within a nxn confusion matrix M for a given element α and given that columns are predicted values and rows are actual values. The diagonals are true positives and every other element is a mismatch.

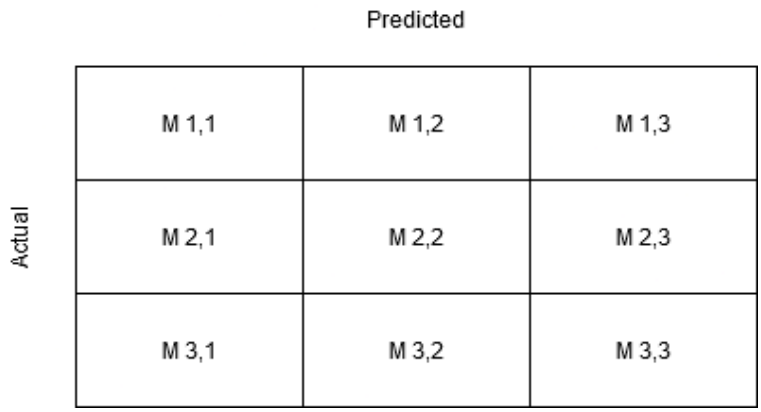

Figure 5 Reference for 3x3 confusion matrix

$$\text{Accuracy} = \frac{\sum_{i=0}^{n} M_{i,i}}{n}$$
$$\text{Precision}_{\alpha} = \frac{M_{\alpha,\alpha}}{\sum_{i=0}^{n} M_{i,\alpha}}$$
$$\text{Recall}_{\alpha} = \frac{M_{\alpha,\alpha}}{\sum_{i=0}^{n} M_{\alpha,i}}$$
$$\text{F1}_{\alpha} = \frac{2\text{Precision}_{\alpha}\text{Recall}_{\alpha}}{\text{Precision}_{\alpha} + \text{Recall}_{\alpha}}$$

Accuracy is the general measurement of the model's ability to label text correctly. Precision of a given label is the amount of data that are correctly classified, recall of a given label is the amount of points classified under that classification. F1 is the mean of both precision and recall.

Model Deployment and Visualization

After the model is evaluated, the model is used for analysis for polarity within the analysis dataset (All The News 2.0) in order to measure the polarity of articles within various news publications. Prior to deployment, the dataset must also be cleaned in a similar manner to the data cleaning section as to ensure accurate results. The dataset is also labelled via AllSides (*AllSides*, n.d.) in order to categorize them by

media bias labels ranging from far left to far right. The resulting statistics will be noted and published in the form of various graphs such as a scatter plot or bar chart which is frequently used for visualizing multiple dimensions of data at once. These visualizations will be made using matplotlib and the results of the analysis will be meticulously interpreted.

References

Allsides. (n.d.). Retrieved 21-12-2024, from https://www.allsides.com

Baccianella, S., Esuli, A., Sebastiani, F., et al. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec* (Vol. 10, pp. 2200–2204).

David D. Lewis, T. G. R., Yiming Yang, & Li, F. (2004). *Rcv1: A new benchmark collection for text categorization research* (Vol. 5). Retrieved from http://www.jmlr.org/papers/volume5/lewis04a/lewis04a.pdf

Ferretti, J. (n.d.). *Bbc articles dataset with extra features.* Retrieved 21-12-2024, from https://www.kaggle.com/datasets/jacopoferretti/bbc-articles-dataset/data

Mu, G., Li, J., Li, X., Chen, C., Ju, X., & Dai, J. (2024). An Enhanced IDBO-CNN-BiLSTM Model for Sentiment Analysis of Natural Disaster Tweets. *Biomimetics*, *9*(9). (Publisher: Multidisciplinary Digital Publishing Institute (MDPI))

Preda, G. (n.d.). *Bbc news.* Retrieved 21-12-2024, from https://www.kaggle.com/datasets/jacopoferretti/bbc-articles-dataset/data

Ravi, K., & Vela, A. E. (2024). *Comprehensive dataset of user-submitted articles with ideological and extreme bias from reddit.* (Vol. 56). Elsevier.

Sebastiani, F., & Esuli, A. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th international conference on language resources and evaluation* (pp. 417–422).

Singh, A. (n.d.). *Nyt articles: 2.1m+ (2000-present) daily updated.* Retrieved 21-12-2024, from https://www.kaggle.com/datasets/aryansingh0909/nyt-articles-21m-2000-present

Thompson, A. (n.d.). *All the news 2.0.* Retrieved 21-12-2024, from https://components.one/datasets/all-the-news-2-news-articles-dataset/

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, *55*(7), 5731–5780.