Optimisation of Good News Classification using Large Language Models

Hannah M. Claus
School of Electronic Engineering and Computer Science
Queen Mary University of London
London, United Kingdom
h.m.claus@se22.qmul.ac.uk

***Abstract*—This research paper comprehensively explores the performance and implications of various machine learning models in sentiment-aware news classification. Through detailed analyses of accuracy, loss, and model comparisons, the study showcases the strengths and limitations of diverse architectures, ranging from traditional classifiers, such as CNNs and RNNs, to advanced transformer-based models, such as DistilBERT. Ethical considerations and potential biases within these models are also meticulously examined.**

***Index Terms*—news, optimisation, NLP, text classification**

I. Introduction

In today's digital age, news consumption plays a crucial role in shaping public opinion and influencing societal well-being. However, the prevalence of negative news content has been associated with adverse effects on people's mental health and overall perception of the world (Sufi 2022). To address this issue, this thesis aims to develop a variety of machine learning (ML) classifiers capable of discerning between positive and negative news. The developed classifiers will be optimised and compared against existing large language models (LLMs) to provide a practical framework to be implemented to provide people with uplifting news content, potentially improving their mental health and supporting positive social progress.

A. Motivation

Currently, the task of curating uplifting news content relies heavily on trained journalists who manually sift through numerous news providers to identify positive stories (Network 2023). For instance, after an interview with the founder of the Good News Network, Geri Weis-Corbley, it became evident, that her main approach to choosing what is a 'good' news article is by thinking about how it would make readers feel (ibid.). This manual selection process, although effective, can be time-consuming, resource-intensive, and strongly subjective. To streamline and automate this process, the proposed approach leverages the efficiency of neural networks to develop a complex classifier capable of automatically identifying positive news headlines using natural language processing (NLP).

B. Research Question

This thesis focuses mainly on developing and optimising a sentiment-aware news classifier as a framework and optimising an already existing LLM. The purpose of this classifier is to assist news providers and news consumers in filtering for positive news and ultimately improve their mental health.

This work will find possible solutions to the following questions:

- How does the classifier learn to recognise different subjective sentiments?
- How do different optimisation techniques affect the performance of the developed classifier compared to the baseline model?
- How does the developed classifier compare to existing LLMs?
- What are the ethical considerations associated with curating news content using ML algorithms, and how can potential biases and limitations be addressed in the development and deployment of the proposed classifier?

This can be achieved by following a few important steps as depicted in Figure 1. These steps create a clear structure for the successful development of the final framework.

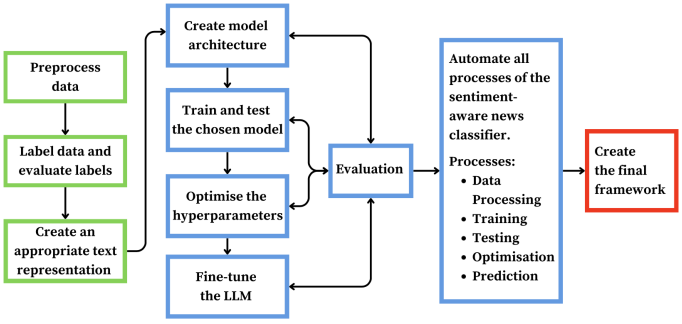

Figure 1. Overview of the main steps of the framework.

II. Related Work

The theoretical framework of this thesis draws upon several key areas of ML, including NLP, sentiment analysis (SA), and LLMs. These frameworks provide the foundation for understanding and implementing the ML techniques necessary to develop the desired classifier.

A. Theoretical Framework

By integrating the following theoretical constructs, this work aims to contribute to the advancement of news curation and the promotion of positive content dissemination in society.

1) Natural Language Processing: NLP encompasses a set of computational techniques and methodologies that enable computers to understand and process human language (Cambria and White 2014). NLP techniques play a crucial role by facilitating the analysis and classification of news based on their textual content. By utilising NLP algorithms, features such as word frequency, semantic meaning, and syntactic structures can be extracted from news headlines and articles, allowing for the creation of meaningful representations that can be further used for SA.

2) Sentiment Analysis: SA, also known as opinion mining, focuses on determining the sentiment or emotional tone expressed in a piece of text (Pang and Lee 2008). In the context of this thesis, SA techniques will be employed to identify and classify the sentiment conveyed in news headlines and articles as either positive or negative. This analysis will enable the development of a classifier that can automatically distinguish between uplifting and discouraging news content. SA algorithms can leverage various approaches, including lexicon-based methods, ML, and deep learning (DL), to capture and interpret the sentiment of textual data.

3) Large Language Models: LLMs have emerged as powerful tools in natural language understanding and generation tasks (Brown et al. 2020). LLMs, such as OpenAI's GPT-3 (Generative Pre-trained Transformer 3), are pre-trained on vast amounts of text data and can generate coherent and contextually relevant text. Here, LLMs are utilised to compare and optimise the performance of the developed classifier against existing state-of-the-art models. By exploring the capabilities of LLMs, insights can be gained into the potential enhancements in the accuracy and efficiency of the classifier. Specifically, for this thesis, a BERT model (Bidirectional Encoder Representations from Transformers) is chosen. It is a pre-trained NLP model developed by Google AI (Devlin et al. 2019). It is designed to capture contextual information from both directions (left-to-right and right-to-left) of a text, unlike previous models that only considered one direction. BERT's architecture is based on the Transformer model, utilising attention mechanisms to weigh the importance of different words in a sentence based on their relationships with each other (Vaswani et al. 2017). This bidirectional context understanding enables BERT to excel in various NLP tasks, such as language understanding, SA, and question answering, without requiring task-specific architecture modifications (Devlin et al. 2019). A distilled version of the BERT base model, DistilBERT, is implemented as it is smaller, faster, cheaper and lighter, and yet keeps 97% of its language understanding capabilities (Sanh et al. 2020).

B. Literature Review

News classification has been a topic of interest in the field of information retrieval and NLP. Researchers have explored various approaches to categorise news headlines based on different criteria such as topic, sentiment, and credibility. For instance, Severyn and Moschitti (2016) proposed a supervised learning framework for news topic classification using convolutional neural networks (CNNs). Their work demonstrated the effectiveness of DL techniques in accurately categorising news headlines into specific topics. While topic classification is valuable, the focus of this paper is to discriminate between positive and negative news, providing a novel contribution to the field.

SA, particularly in the context of news content, has gained significant attention due to its potential applications in understanding public opinion, sentiment trends, and content recommendation systems. For instance, Patodkar and I.R. (2016) conducted a comprehensive survey of SA techniques and highlighted the challenges in accurately determining sentiment in news headlines. Their survey serves as a valuable reference for understanding the nuances and complexities associated with SA tasks.

The concept of providing uplifting news content to improve mental well-being has garnered attention in recent years. Research suggests that exposure to negative news can have negative effects on individuals' moods, emotions, and overall perception of the world (Johnston and Davey 1997). Additionally, the psychological impact of negative news consumption has been linked to increased stress levels and anxiety (Zhang et al. 2021). By developing a classifier that discriminates between positive and negative news, this thesis aims to contribute to the promotion of good news consumption.

III. Methodology

The following section describes all stages that lead to the development of the framework.

A. Concept

As no universal rule defines whether a subjective statement is inherently positive or negative, objective measures must be established. Thus, a purely threshold-based approach is not enough to capture the overall sentiment of a news headline or article. For this reason, a classifier is trained to recognise specific patterns from news headlines and articles pre-labelled as positive or negative.

B. Dataset

The dataset consists of ca. 143,000 news articles and their headlines from 15 different American news providers between 2016 and 2017 (Thompson 2017). All articles are written in American English and include various topics. The distribution of the publications is shown in Fig. 2.

The articles stretch over 15 months between 2016 and 2017. It is important to mention that 2016 was a very eventful year for the US, as Donald Trump was elected president in November and took over after Barack Obama. When looking at the 20 most frequently used words throughout the entire dataset, this political event had a strong impact on the news coverage. With 'trump' being the most frequent word by far, and connected politicians' names, such as 'clinton' and 'obama' close behind. The words 'president', 'house', and 'white' are also represented, supporting the assumption that the elections and the presidency dominated the American news

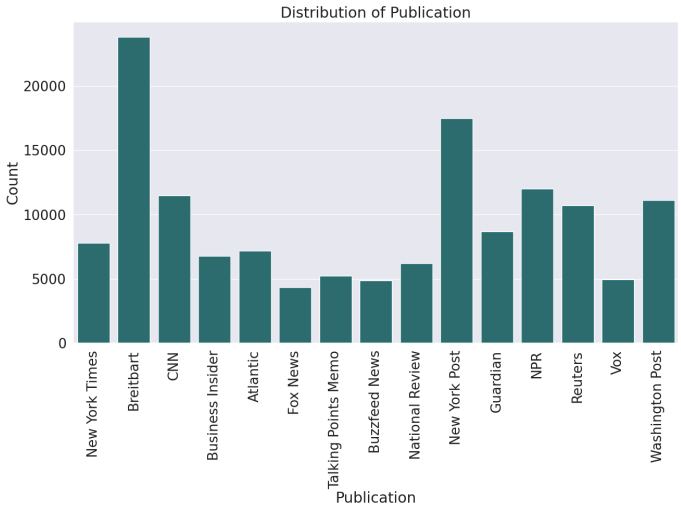

Figure 2. Distribution of different publications.

coverage in 2016 and 2017. The frequency distribution is further depicted in Figure 3.

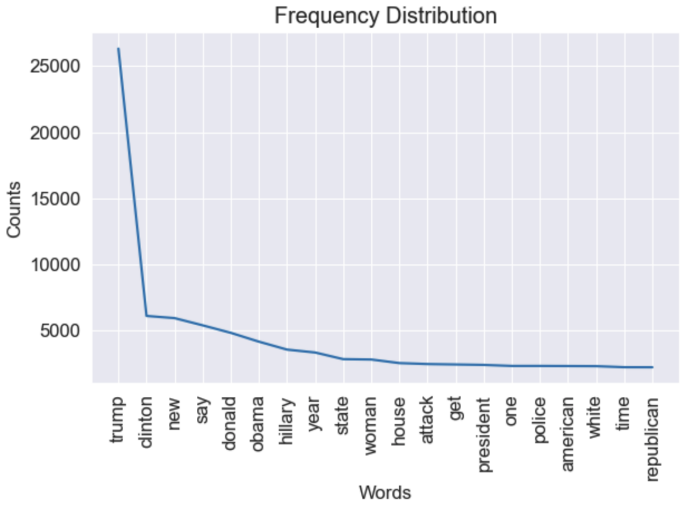

Figure 3. Top 20 most frequent words in the entire dataset.

Furthermore, when creating a Skip-gram model, which is one of the unsupervised learning techniques to find the most related words for a given word (Guthrie et al. 2006), a t-distributed stochastic neighbour embedding (t-SNE) is created to analyse the word similarities within the dataset (Van der Maaten and Hinton 2008). The five coloured circles in Figure 4 point at supported connections from specific events. For instance, the blue circle shows that 'jong' and 'un' appear to be shown in similar contexts, possibly due to North Korea's politician Kim Jong Un, who appeared in the news because of North Korea's missile and nuclear tests in 2016 (Nikitin 2019). However, 'kim' is further away due to its shared connection to the American businesswoman and media personality Kim Kardashiann (Stewart 2015). This is further underlined by analysing the cosine similarities that returned 'jong', 'tsang', 'nam', and 'kardashian' as the four closest words to 'kim'. Moreover, the red circle shows words appearing in close context with 'trump', as former president Trump discussed several plans of his presidential agenda, mentioning 'climate', 'missile' (connected to North Korea again), and further 'form[s]' (Ryan 2021).

Figure 4. T-SNE plot to depict the word similarities generated by a Skip-gram model.

Given the above analysis of the dataset, it becomes evident that the period of the news content included in the dataset is heavily influenced by the American elections in 2016. This leads to an unusually higher appearance of the names of the then-active politicians Trump, Obama, and Clinton. To test the classifiers' robustness, the predictions of labels are conducted by introducing new and unseen data from an Australian news dataset provided by ABC (Australian Broadcasting Corporation) over 19 years (Corporation 2021).

C. Development of the Classifier

The task of the sentiment-aware news classifier is to identify the sentiment of the provided (a) news headline, and later (b) news article. The text is classified as either (1) positive, (0) negative, (-1) or neutral in case it cannot be classified as either positive or negative.

D. Text Preprocessing

After reading the data, the news headlines and articles need to be preprocessed and cleaned to transform the raw text into a structured format that can be efficiently analysed (Perkins 2014).

1) Text Cleaning: The initial step in the preprocessing pipeline involves removing specific publishers' names from the text data (e.g. The New York Times). This step is crucial for eliminating potential bias or unwanted information associated with specific sources.

2) Number and Punctuation Removal: All numerical and punctuation characters are removed from the text. This step removes the presence of these characters in the text, which may not contribute significantly to certain NLP tasks.

3) Tokenisation: The text is split into individual tokens, resulting in a list of separate tokens. All tokens are then converted to lowercase to ensure case insensitivity and reduce the vocabulary size.

4) Stop Word Removal and Lemmatisation: Stop words, such as commonly used words like "the," "is," and "and", are removed from the tokens to reduce noise and focus on content-specific words. Consecutively, lemmatisation is applied to the remaining tokens to reduce inflected words to their base or dictionary form. This helps in achieving better feature representation by reducing morphological variations.

The final output of the text preprocessing is a list of processed tokens that are used as input data for the classifiers.

E. Data Labelling

Once the text is fully preprocessed, the news headlines and articles are labelled for training and testing using automated labelling techniques. Here, four different methods are used. Each news headline and article receives four labels in the end, so the average label is taken to create one general label for training purposes.

1) Snorkel: Snorkel is a programmatically assisted labelling technique that aims to automate the generation of training data for ML models (Ratner et al. 2017). It leverages weak supervision to create large labelled datasets using heuristics or labelling functions. These labelling functions are developed based on domain knowledge or rule-based approaches, allowing the system to automatically assign labels to unlabelled data points (ibid.).

2) TextBlob: TextBlob is a Python library that provides a simplified interface for performing common NLP tasks, including text classification and SA (Loria n.d.). TextBlob offers a pre-trained SA model that utilises a lexicon-based approach. It assigns sentiment labels, such as positive, negative, or neutral, to text based on the presence and polarity of words in the provided text. TextBlob's SA is a straightforward and accessible technique that does not require extensive training or complex models (ibid.).

3) Dictionary-based labelling: Dictionary-based labelling is a technique commonly used in SA to classify text based on the presence or absence of specific words or phrases associated with positive or negative sentiment (Liu and L. Zhang 2012). This approach involves constructing a sentiment lexicon or dictionary containing predefined words or phrases with known sentiment polarities. Each word or phrase in the text is matched against the lexicon, and the sentiment label is assigned based on the presence and sentiment polarity of the matching terms. Dictionary-based labelling provides a simple and interpretable method for sentiment classification but may overlook contextual nuances and sarcasm (ibid.).

4) VADER: VADER (Valence Aware Dictionary and sEntiment Reasoner) is a rule-based sentiment analysis tool specifically designed for social media text (Hutto and Gilbert 2014). It employs a combination of lexical and grammatical heuristics to determine sentiment polarity in text. VADER utilises a pre-constructed sentiment lexicon that contains a vast array of words, each assigned with intensity scores representing their sentiment polarity. By considering word order, capitalisation, punctuation, and other contextual features, VADER computes a sentiment score for the text, indicating the overall positive, negative, or neutral sentiment. VADER is particularly useful for analysing informal and colloquial language commonly found in social media texts (ibid.).

F. The Neural Network

Text classification is a fundamental task in NLP that involves categorising text documents into predefined classes (Kowsari et al. 2019). In this context, the goal is to determine whether a given news headline or article expresses an overall positive or negative sentiment. This part outlines the models and text representations used to perform the task.

1) Text Representation: To represent the text data in a numerical format suitable for ML algorithms, different text representation techniques are employed: (a) TF-IDF (Term Frequency-Inverse Document Frequency) is a widely used method that reflects the importance of a word in a document or a corpus. It assigns a weight to each word based on its frequency in the document and the inverse frequency across all documents (Wang and Manning 2012). (b) CountVectorizer represents text documents as vectors of word counts. It creates a feature vector by counting the occurrences of each word in the document (Goyal 2021). (c) Word2Vec is a word embedding technique that represents words as dense vectors in a continuous vector space. It captures semantic relationships between words by training a neural network on a large corpus of text (Mikolov et al. 2013).

2) Models: The classifier includes various models built using the Scikit-Learn (SK) and Tensorflow (TF) libraries for this task:

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- Gaussian Naive Bayes (GNB)
- Basic baseline model (Basic)
- Convolutional Neural Network (CNN)
- Long Short-Term Memory (LSTM)
- Bidirectional LSTM (BiLSTM)
- Recurrent Neural Network (RNN)
- DistilBERT

Both libraries implement different architectures. To generalise, the models will be split by the type of word embedding used and referenced as being either an SK model or a TF model. Figure 5 depicts the general architecture of an SK model, which includes LR, SVM, and GNB.

These following are widely used in text classification tasks due to their simplicity, effectiveness, and interoperability (Minaee et al. 2021). The LR model is a linear model that predicts the probability of a class label based on the input features (Mikolov et al. 2013). The SVM model is a discriminative model that constructs a hyperplane to separate different classes (Veropoulos, Campbell, Cristianini, et al. 1999). The GNB model assumes that features are conditionally independent given the class label and calculates the likelihood of each class using Gaussian distributions (Bird, Klein, and Loper 2009).

In contrast, the TF models implement an embedding layer that maps from integer indices, which stand for specific

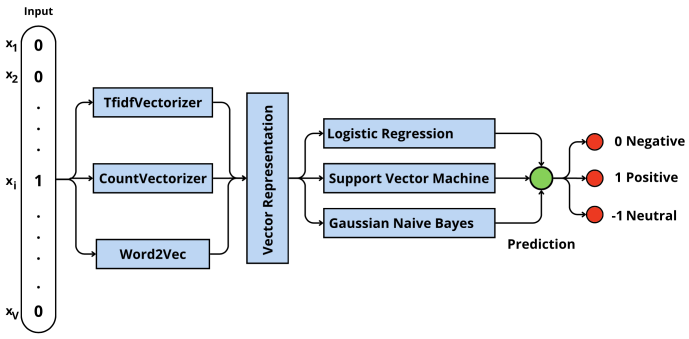

Figure 5. Architecture of an SK model.

words, to dense vectors, which represent their embeddings (TensorFlow 2023).

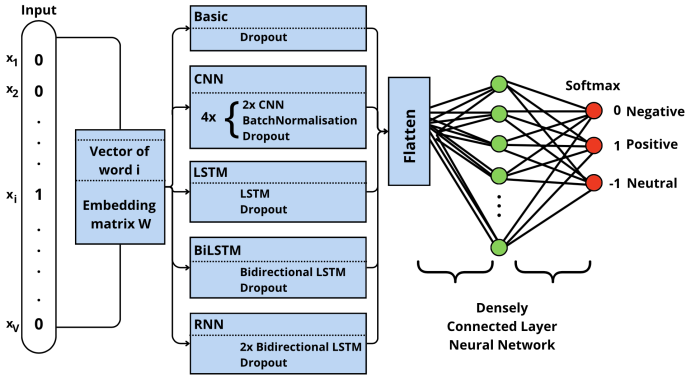

Figure 6. Architecture of a TF model.

As shown in Figure 6, five different models are implementing the embedding layer. The Basic model consists of the embedding layer and a dropout, to create a simple model. The structure of the Basic model is used for the rest of the models, by extending it with specific layers. The CNN for instance adds 8 convolutional layers. CNNs add convolutional layers to the network by using kernels. The kernels stride along the input matrix and a feature map is generated. This way the data's complexity is captured in a way that includes spatial dependencies as well (Jacovi, Shalom, and Goldberg 2020). To include some context between the word embeddings, an LSTM is added. LSTMs have feedback connections which make them different from more traditional feed-forward neural networks. Using this, LSTMs process entire sequences of data without treating each point in the sequence independently. Instead, LSTMs retain useful information about previous data in the sequence to help with the processing of new data points (Sherstinsky 2020). The LSTM model is then extended to include first one BiLSTM layer in the BiLSTM model, and then two BiLSTM layers in the RNN model, as now the first BiLSTM layer returns the learned sequences for the next layer to incorporate the new information. The BiLSTM is a sequence processing model that consists of two separate LSTMs. One LSTM takes the input in a forward direction, and the other LSTM takes it in a backward direction. This way, the amount of information available to the network is effectively increased, improving the context available to the algorithm (Hochreiter and Schmidhuber 1997).

Lastly, the aforementioned LLM called DistilBERT is fine-tuned for this specific task to compare the performance to the SK and TF models. The fine-tuning process also includes various data augmentation methods, such as Synonym Word Augmentation (SWA), which substitutes words by leveraging semantic meaning, and Random Word Deletion (RWD), which deletes random words in the data sequence (Wei and Zou 2019). Another method tokenises the words using a text tokeniser that is specifically geared towards social networks. In this case, a Twitter tokeniser (TW) is used, as some of the news headlines have similar lengths and structures as general Twitter posts (Baziotis, Pelekis, and Doulkeridis 2017).

G. Optimisation

Furthermore, this section introduces four optimisation methods: Hyperband (HB), Random Search (RS), and Bayesian optimisation (BO), which enhance the model performance by tuning the hyperparameters. For this thesis, the optimisation methods are only applied to the TF models due to their simple integration and efficiency.

1) Hyperband: Hyperband is a hyperparameter optimisation method that combines multi-armed bandit strategies with a successive halving algorithm. It efficiently explores the hyperparameter space by allocating resources to promising hyperparameter configurations while discarding unpromising ones. This process leads to improved model performance with limited computational resources (Li et al. 2018).

2) Random Search: Random Search is a simple optimisation method that randomly samples hyperparameter configurations from a predefined search space. By exploring different regions of the search space, it has the potential to discover good hyperparameter settings (Bergstra and Bengio 2012). Although it may require more iterations to converge, it is less computationally expensive compared to grid search (Claus 2022).

3) Bayesian Optimisation: Bayesian optimisation is a sequential model-based optimisation method that uses Bayesian inference to find the global optimum of a black-box function. It models the unknown function with a probabilistic surrogate model and iteratively selects the next hyperparameter configuration based on an acquisition function that balances exploration and exploitation. Bayesian optimisation has shown promising results in hyperparameter optimisation, especially for expensive-to-evaluate functions (Snoek, Larochelle, and Adams 2012).

IV. Evaluation

The process of training, testing, and optimising the various models has been fully automated and put together in a framework. This framework offers the tools needed to apply the sentiment-aware news classifier to any given dataset and reproduce the following results.

A. Results

The main focus of the training, testing, and optimisation lies in processing the news headlines of the dataset. Due to their manageable size, it is more efficient to start the experiments with the headlines and use the gathered information on the development of a classifier focused on news articles. Thus, the following five tables show accuracies received by using the news headlines dataset only and the last table shows the results using entire news articles. Thus, the experiments are divided into six parts for a better overview. The effectiveness of each model is evaluated based on accuracy and loss metrics, shedding light on their predictive capabilities. The tables provide a comprehensive overview of the performance of different models, facilitating a thorough comparison and understanding of their strengths and weaknesses.

First, the SK models are trained and tested. As can be seen in Table I, all three models are trained with different vectorisers. However, as soon as the dataset reached more than 10,000 entries, the W2V models stopped working, hinting at the complexity of the models. Even the LR W2V model performs poorly, indicating that W2V might not be well-suited for this task. However, the other LR models exhibit moderate accuracy scores, with the best results achieved by LR CV at 70.6% in testing. SVM models, particularly SVM TF-IDF, attain a training accuracy of 76.8%, although their testing accuracy is comparatively lower. Gaussian Naive Bayes (GNB) models perform relatively worse, seeming to be underperforming for this task.

	Accuracy	
Model	Training	Test
LR TF-IDF	71.0%	70.4%
LR CV	71.1%	**70.6%**
LR W2V	43.4%	36.0%
SVM TF-IDF	**76.8%**	65.6%
SVM CV	70.4%	70.0%
GNB TF-IDF	65.2%	64.3%
GNB CV	64.8%	64.0%

Table I
PERFORMANCE OF THE SKLEARN CLASSIFIERS

In Table II, the focus shifts to the performance of the TF classifiers. The basic model exhibits poor predictive ability, achieving only 30.0% training accuracy and 29.7% test accuracy. This indicates that the model is not complex enough to capture the complexities of the sentiment in news headlines. More advanced models, such as CNN, LSTM, BiLSTM, and RNN, demonstrate significantly improved results. Unfortunately, the CNN model seems to suffer from overfitting as evidenced by the significantly lower test accuracy. Among these other models, the RNN model stands out with the highest accuracy of 88.0% in training and 86.0% in testing, accompanied by a low loss of 0.40. This suggests that the RNN architecture is best suited for this task out of all the SK and TF models.

	Accuracy		
Model	Training	Test	Loss
Basic	30.0%	29.7%	-
CNN	87.8%	54.2%	1.27
LSTM	86.5%	85.5%	0.41
BiLSTM	87.9%	85.9%	0.41
RNN	**88.0%**	**86.0%**	**0.40**

Table II
PERFORMANCE OF THE TENSORFLOW CLASSIFIERS

After establishing the base accuracy for the above models, the three optimisation methods are implemented to search the hyperparameter space for the optimal values. As three of the four complex TF models (CNN, LSTM, BiLSTM, and RNN) utilise at least one LSTM, the optimisation is fully focused on optimising the LSTM model to choose the best optimisation technique to be applied to the other models. As depicted in Table III, BO, HB, and RS are applied to the LSTM model first. Even though LSTM with BO achieves the highest training accuracy thus far, it also has a greater distance to the test accuracy and one of the highest losses. Additionally, it lowers the testing accuracy from the previously achieved 85.5% to 82.9%. LSTM RS and LSTM HB perform well, with LSTM HB having a slightly higher test accuracy, lower loss, and smaller gap between the training accuracy. Hence, HB is chosen as the optimisation method to implement for the other three models. When applying HB to the BiLSTM model, it slightly improves the training accuracy and loss, however, it does not change the test accuracy. Similarly, the RNN model does not improve much, it appears that the hyperparameter search does not manage to find better parameter values than the base model, where the values are manually chosen. The only significant change is seen in the CNN model, whose test accuracy improves by almost 30%. Nonetheless, the training accuracy is still too high for the test accuracy, which shows that overfitting is still happening. As a result, BiLSTM HB performs the best among the optimised TF models, achieving both high training and test accuracy, as well as low loss.

	Accuracy		
Model	Training	Test	Loss
LSTM BO	**95.4%**	82.9%	0.86
LSTM RS	91.6%	85.6%	0.41
LSTM HB	90.3%	**85.7%**	**0.39**
CNN HB	**97.5%**	83.7%	0.87
BiLSTM HB	90.8%	**85.9%**	0.40
RNN HB	90.7%	85.4%	0.42

Table III
PERFORMANCE OF THE OPTIMISED CLASSIFIERS

After training, testing, and optimising the above models, DistilBERT is implemented. As it is a pre-trained model specifically geared toward learning the inner representation of the English language to be fine-tuned on a downstream task, it is an appropriate model for this task (Sanh et al. 2020).

Table IV shows that the base DistilBERT model returns good accuracy on both training and test data, along with moderate loss values. DistilBERT TW, the version of DistilBERT that processes the text by implementing a Twitter tokeniser, performs the best, with the highest accuracy on both training and test data, as well as the lowest loss value. Adding first SWA, then RWD and finally both methods together seems to decrease the model's accuracy, indicating that this modification might not be beneficial for this task.

	Accuracy		
Model	Training	Test	Loss
DistilBERT	87.5%	85.3%	0.59
DistilBERT TW	**88.1%**	**87.5%**	**0.38**
DistilBERT SWA	79.7%	74.4%	0.68
DistilBERT RWD	70.0%	70.4%	0.74
DistilBERT Both	65.0%	61.9%	0.89

Table IV
PERFORMANCE OF THE DISTILBERT CLASSIFIERS

Table V summarises the performance of the best classifiers from the previous tables. RNN, DistilBERT TW, and BiLSTM HB emerge as top performers, exhibiting competitive accuracy and loss values. It becomes evident that using the TF models that include at least two LSTMs is important to achieve good results.

	Accuracy		
Model	Training	Test	Loss
RNN	88.0%	86.0%	0.40
LR CV	71.1%	70.6%	-
DistilBERT TW	88.1%	**87.5%**	**0.38**
BiLSTM HB	**90.8%**	85.9%	0.40

Table V
SUMMARY OF THE BEST CLASSIFIERS

Following the results of the experiments using the news headlines dataset only, the RNN and BiLSTM HB architectures are implemented to train and test with news articles. The results shown in Table VI show that both models achieve moderately good results, with the RNN model slightly outperforming the BiLSTM model with a test accuracy of 77.3%.

	Accuracy		
Model	Training	Test	Loss
RNN	**80.0%**	**77.3%**	0.54
BiLSTM HB	79.9%	77.1%	0.54

Table VI
TRAINING WITH NEWS ARTICLES

All in all, a total amount of 25 different models have been trained, tested, and optimised on two different datasets using the framework to create a sentiment-aware news classifier. In the end, the RNN model consisting of two bidirectional LSTM layers achieves the best accuracy with 86.0%. Moreover, after fine-tuning the DistilBERT model, it slightly outperforms the RNN model with 87.5% test accuracy.

B. Analysis

Sentiment-aware text classification plays a pivotal role in extracting subjective information from text data, enabling a better understanding of public opinions and emotions, especially in news content. In this section, the results obtained from a comprehensive evaluation of different ML models are analysed. The analysis encompasses the models' performance, their strengths and limitations, as well as the implications of these findings in the context of sentiment-aware classification of news content. The first noteworthy observation emerges from the comparison of the TF models in Table II. The basic model's dismal accuracy underscores the complexity of sentiments in news headlines. However, the subsequent models - LSTM, BiLSTM, and RNN - demonstrate substantial performance improvements, emphasising their ability to capture temporal dependencies inherent in sequences of words. This stresses the significance of recurrent structures, such as LSTMs in SA tasks, where understanding the sequential nature of language proves crucial. In contrast, the SK models demonstrate limited success, likely due to the underlying assumption of feature independence, which may not hold for complex language data. Collectively, these results reinforce the notion that DL approaches, particularly RNNs, surpass traditional ML models in capturing intricate patterns within text data. Furthermore, CNN HB and LSTM HB further reinforce the significance of hyperparameter optimisation, providing a delicate balance between accuracy and computational efficiency. These results highlight the potential of hyperparameter optimisation in refining models to achieve better generalisation and performance. However, as BiLSTM HB and RNN HB show, the use of these optimisation methods does not work every time and may cause unnecessary overfitting. Thus, the results of the model need to be properly analysed before using the model in real-world applications. Then again, the DistilBERT models exhibit competitive accuracy levels. DistilBERT TW emerges as a standout performer, surpassing other models in both training and testing accuracy, while maintaining a remarkably low loss. This suggests that transformer architectures, with their attention mechanisms and contextual embeddings, excel at grasping nuanced semantics and sentiment nuances within news headlines. The success of these models underscores the paradigm shift that transformer architectures have brought to various NLP tasks (Vaswani et al. 2017). Nevertheless, when using news articles instead of news headlines both RNN HB and BiLSTM HB exhibit resilience in their predictive capabilities, maintaining competitive accuracy levels when trained on longer and more complex text inputs. This observation suggests that the selected models are capable of scaling their performance to handle more extensive and intricate textual content, thereby widening their application scope beyond short headlines.

V. DISCUSSION

This discussion section delves into the broader implications of the above findings, considering the presence of biases within

the models and the ethical considerations associated with their deployment.

A. Ethical Implications

While the models exhibited promising accuracy levels, it is imperative to acknowledge the potential biases inherent in their predictions. Biases can emerge from the training data and impact the models' ability to generalise accurately. Bias mitigation techniques, such as debiasing training data and carefully selecting diverse data sources, can be employed to counteract this challenge. As described during the initial analysis of the dataset, the word frequencies are heavily influenced by the American elections of 2016. Thus, applying the model to American-centered or Europe-centered news content will most likely have a better accuracy than African news content for instance, due to different political agendas. Additionally, the names of politicians might not carry the same importance in 10 years, showing that a classifier for news content needs to be constantly updated due to the rapid pace of change in society. Moreover, biases can emerge from the language itself. Certain phrases, words, or idioms might be incorrectly classified as positive or negative due to cultural nuances or historical context. Models that generalise based on sentiment patterns within a specific dataset might struggle when exposed to different language styles or sentiments. Addressing these challenges requires continuous monitoring, feedback loops, and ongoing model updates to adapt to changing language and societal norms. Important to mention is the potential for misclassification due to a repetition of names, as already demonstrated with the word similarities of 'kim'. Kim Jong Un will most likely appear in more negative news content due to the tense political relationship with North Korea (Nikitin 2019), whereas Kim Kardashian often appears in entertainment-related content that is mostly geared towards positive news content (Stewart 2015). This way, even a person's name can have ambiguity, which needs to be properly mitigated. In general, the deployment of sentiment-aware models in real-world scenarios carries ethical responsibilities. News headlines significantly influence public perceptions, and incorrect sentiment predictions could inadvertently contribute to misinformation, anxiety, or the amplification of biased narratives. Especially when dealing with improving people's mental health by recommending more positive news content, errors can lead to health-related issues, such as stress or depression (Zhang et al. 2021). Additionally, ethical considerations extend to data privacy and consent. The models' performance is contingent upon vast amounts of training data, potentially comprising user-generated content. Ensuring that data used for training is collected ethically, with explicit consent and data anonymisation, is paramount to respecting user privacy and data rights. In this case, the dataset is chosen due to its simple public access and does not incorporate any confidential user data. However, ensuring fairness in these models is essential to prevent biased outcomes and discrimination. Model outputs should not disproportionately favour or disadvantage any particular group. Buolamwini's work emphasises the need for vigilance in designing and evaluating ML models to prevent unintended biases, ensuring fair treatment of diverse viewpoints (Buolamwini 2017).

B. Psychological Impacts

The psychological impact of negative news is a critical aspect beyond the technical and ethical realms. Research shows that exposure to negative news, for instance during the COVID-19 pandemic, significantly affects mental well-being and perspectives (Zhang et al. 2021). Such exposure correlates with heightened stress, anxiety, and feelings of helplessness. In contrast, positive or balanced news content enhances mood and well-being, fostering optimism and hope (McIntyre and Gibson 2016). As the developed models categorise news content, they carry the responsibility of shaping emotional responses and mental states (Wild et al. 2019). Hence, awareness and management of psychological effects are essential, ensuring a balanced and healthy news consumption experience that can be personalised through user feedback.

VI. Conclusion

All things considered, the outcomes of this thesis have several implications for the field of sentiment-aware text classification as further described in the following.

A. Summary

Firstly, the research emphasises the importance of selecting appropriate model architectures based on the nature of the data and the problem. Secondly, it underscores the ascendancy of transformer-based models in capturing context and semantics effectively. Furthermore, the study demonstrates the impact of architecture optimisation and hyperparameter tuning in enhancing predictive accuracy while maintaining computational efficiency. Lastly, the research highlights the potential of adapting models to diverse text inputs, paving the way for more comprehensive sentiment analysis in real-world scenarios.

B. Outlook

In future research, exploring hybrid models that combine the strengths of different architectures, leveraging ensemble techniques, and investigating transfer learning approaches may yield further advancements. Additionally, addressing class imbalance, analysing misclassified instances, and experimenting with domain-specific pre-training could enhance model robustness and real-world applicability, for instance by creating international datasets that focus less on Western-oriented news. For user-oriented applications, the models could be integrated into a user interface that supports user validation and feedback, allowing personalised recommendations of news content. The findings presented in this thesis offer valuable insights into the dynamic landscape of sentiment analysis, fostering innovation and progress in this evolving field. This way, good news consumption is made easier and news-related, potentially decreasing stress and anxiety.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr Bruno Ordozgoiti at Queen Mary University of London for his continuous support and motivation during my Master's Thesis, and for always believing in my research.

References

Baziotis, Christos, Nikos Pelekis, and Christos Doulkeridis (Aug. 2017). "DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis". In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, pp. 747–754. DOI: 10.18653/v1/S17-2126. URL: https://aclanthology.org/S17-2126 (visited on 08/24/2023).

Bergstra, James and Yoshua Bengio (2012). "Random Search for Hyper-Parameter Optimization". en. In.

Bird, Steven, Ewan Klein, and Edward Loper (June 2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. en. Google-Books-ID: KGIbfiiP1i4C. "O'Reilly Media, Inc." ISBN: 978-0-596-55571-9.

Brown, Tom et al. (2020). "Language Models are Few-Shot Learners". en. In: *Advances in Neural Information Processing Systems* 33, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf (visited on 08/24/2023).

Buolamwini, Joy Adowaa (2017). "Gender shades : intersectional phenotypic and demographic evaluation of face datasets and gender classifiers". eng. Accepted: 2018-03-12T19:28:30Z ISBN: 9781026503584. Thesis. Massachusetts Institute of Technology. URL: https://dspace.mit.edu/handle/1721.1/114068 (visited on 08/24/2023).

Cambria, Erik and Bebo White (May 2014). "Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]". In: *IEEE Computational Intelligence Magazine* 9.2. Conference Name: IEEE Computational Intelligence Magazine, pp. 48–57. ISSN: 1556-6048. DOI: 10.1109/MCI.2014.2307227.

Claus, Hannah M. (June 2022). "The Importance of Hyperparameter Optimisation for Facial Recognition Applications". en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.11. Number: 11, pp. 13130–13131. ISSN: 2374-3468. DOI: 10.1609/aaai.v36i11.21701. URL: https://ojs.aaai.org/index.php/AAAI/article/view/21701 (visited on 08/24/2023).

Corporation, Australian Broadcasting (2021). *A Million News Headlines*. en. URL: https://www.kaggle.com/datasets/therohk/million-headlines (visited on 08/24/2023).

Devlin, Jacob et al. (May 2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805 [cs]. URL: http://arxiv.org/abs/1810.04805 (visited on 08/23/2023).

Goyal, Radhika (Oct. 2021). "Evaluation of rule-based, CountVectorizer, and Word2Vec machine learning models for tweet analysis to improve disaster relief". In: *2021 IEEE Global Humanitarian Technology Conference (GHTC)*. ISSN: 2377-6919, pp. 16–19. DOI: 10.1109/GHTC53159.2021.9612486.

Guthrie, David et al. (2006). "A Closer Look at Skip-gram Modelling". en. In.

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: *Neural Computation* 9.8. Conference Name: Neural Computation, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735.

Hutto, C. and Eric Gilbert (May 2014). "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text". en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 8.1. Number: 1, pp. 216–225. ISSN: 2334-0770. DOI: 10.1609/icwsm.v8i1.14550. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550 (visited on 08/24/2023).

Jacovi, Alon, Oren Sar Shalom, and Yoav Goldberg (Apr. 2020). *Understanding Convolutional Neural Networks for Text Classification*. arXiv:1809.08037 [cs]. DOI: 10.48550/arXiv.1809.08037. URL: http://arxiv.org/abs/1809.08037 (visited on 08/24/2023).

Johnston, Wendy M. and Graham C. L. Davey (1997). "The psychological impact of negative TV news bulletins: The catastrophizing of personal worries". en. In: *British Journal of Psychology* 88.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8295.1997.tb02622.x, pp. 85–91. ISSN: 2044-8295. DOI: 10.1111/j.2044-8295.1997.tb02622.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-8295.1997.tb02622.x (visited on 08/24/2023).

Kowsari, Kamran et al. (2019). "Text Classification Algorithms: A Survey". In: *Information* 10.4. ISSN: 2078-2489. DOI: 10.3390/info10040150. URL: https://www.mdpi.com/2078-2489/10/4/150.

Li, Lisha et al. (2018). "Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization". en. In.

Liu, Bing and Lei Zhang (2012). "A Survey of Opinion Mining and Sentiment Analysis". en. In: *Mining Text Data*. Ed. by Charu C. Aggarwal and ChengXiang Zhai. Boston, MA: Springer US, pp. 415–463. ISBN: 978-1-4614-3223-4. DOI: 10.1007/978-1-4614-3223-4_13. URL: https://doi.org/10.1007/978-1-4614-3223-4_13 (visited on 08/24/2023).

Loria, Steven (n.d.). "textblob Documentation". en. In: ().

McIntyre, Karen Elizabeth and Rhonda Gibson (Oct. 2016). "Positive News Makes Readers Feel Good: A "Silver-Lining" Approach to Negative News Can Attract Audiences". In: *Southern Communication Journal* 81.5. Publisher: Routledge _eprint: https://doi.org/10.1080/1041794X.2016.1171892, pp. 304–315. ISSN: 1041-794X. DOI: 10.1080/1041794X.2016.1171892. URL: https://doi.org/10.1080/1041794X.2016.1171892 (visited on 08/24/2023).

Mikolov, Tomas et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html (visited on 08/24/2023).

Minaee, Shervin et al. (Apr. 2021). "Deep Learning–based Text Classification: A Comprehensive Review". In: *ACM Computing Surveys* 54.3, 62:1–62:40. ISSN: 0360-0300. DOI: 10.1145/3439726. URL: https://dl.acm.org/doi/10.1145/3439726 (visited on 08/24/2023).

Network, Good News (2023). *GNN Founder Talks With BBC World Service About Positive News in the Media Landscape (Listen)*. URL: https://www.goodnewsnetwork.org/gnn-founder-talks-with-bbc-world-service-about-positive-news-in-media-landscape-listen/ (visited on 08/24/2023).

Nikitin, Mary Beth D (2019). "North Korea's Nuclear and Ballistic Missile Programs". en. In: *North Korea*.

Pang, Bo and Lillian Lee (July 2008). "Opinion Mining and Sentiment Analysis". English. In: *Foundations and Trends® in Information Retrieval* 2.1–2. Publisher: Now Publishers, Inc., pp. 1–135. ISSN: 1554-0669, 1554-0677. DOI: 10.1561/1500000011. URL: https://www.nowpublishers.com/article/Details/INR-011 (visited on 08/24/2023).

Patodkar, V N and Sheikh I.R. (Dec. 2016). "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". en. In: *IJARCCE* 5.12, pp. 320–322. ISSN: 22781021. DOI: 10.17148/IJARCCE.2016.51274. URL: http://ijarcce.com/upload/2016/december-16/IJARCCE%2074.pdf (visited on 08/24/2023).

Perkins, Jacob (Aug. 2014). *Python 3 Text Processing with NLTK 3 Cookbook*. en. Google-Books-ID: rAFcBAAAQBAJ. Packt Publishing Ltd. ISBN: 978-1-78216-786-0.

Ratner, Alexander et al. (Nov. 2017). "Snorkel: Rapid Training Data Creation with Weak Supervision". In: *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases* 11.3, pp. 269–282. ISSN: 2150-8097. DOI: 10.14778/3157794.3157797. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5951191/ (visited on 08/24/2023).

Ryan, Frederick (Dec. 2021). *Opinion — A transcript of Donald Trump's meeting with The Washington Post editorial board*. en-US. URL: https://www.washingtonpost.com/blogs/post-partisan/wp/2016/03/21/a-transcript-of-donald-trumps-meeting-with-the-washington-post-editorial-board/ (visited on 08/24/2023).

Sanh, Victor et al. (Feb. 2020). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv:1910.01108 [cs]. URL: http://arxiv.org/abs/1910.01108 (visited on 08/23/2023).

Severyn, Aliaksei and Alessandro Moschitti (Apr. 2016). *Modeling Relational Information in Question-Answer Pairs with Convolutional Neural Networks*. arXiv:1604.01178 [cs]. DOI: 10.48550/arXiv.1604.01178. URL: http://arxiv.org/abs/1604.01178 (visited on 08/24/2023).

Sherstinsky, Alex (2020). "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network". In: *Physica D: Nonlinear Phenomena* 404, p. 132306. ISSN: 0167-2789. DOI: https://doi.org/10.1016/j.physd.2019.132306. URL: https://www.sciencedirect.com/science/article/pii/S0167278919305974.

Snoek, Jasper, Hugo Larochelle, and Ryan P Adams (2012). "Practical Bayesian Optimization of Machine Learning Algorithms". In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html (visited on 08/24/2023).

Stewart, Martha (Apr. 2015). "Kim Kardashian West: The World's 100 Most Influential People". en-US. In: *Time*. ISSN: 0040-781X. URL: https://time.com/collection-post/3822676/kim-kardashian-west-2015-time-100/ (visited on 08/24/2023).

Sufi, F. K. (2022). "Identifying the drivers of negative news with sentiment, entity and regression analysis". In: *International Journal of Information Management Data Insights* 2.1, p. 100074. ISSN: 2667-0968. DOI: https://doi.org/10.1016/j.jjimei.2022.100074. URL: https://www.sciencedirect.com/science/article/pii/S2667096822000179.

TensorFlow (2023). *Word embeddings — Text — TensorFlow*. URL: https://www.tensorflow.org/text/guide/word_embeddings (visited on 08/23/2023).

Thompson, Andrew (2017). *All the news*. en. URL: https://www.kaggle.com/datasets/snapcrack/all-the-news (visited on 08/24/2023).

Van der Maaten, Laurens and Geoffrey Hinton (2008). "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11.

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (visited on 08/24/2023).

Veropoulos, Konstantinos, Colin Campbell, Nello Cristianini, et al. (1999). "Controlling the sensitivity of support vector machines". In: *Proceedings of the international joint conference on AI*. Vol. 55. Stockholm, p. 60.

Wang, Sida and Christopher Manning (July 2012). "Baselines and Bigrams: Simple, Good Sentiment and Topic Classification". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Jeju Island, Korea: Association for Computational Linguistics, pp. 90–94. URL: https://aclanthology.org/P12-2018 (visited on 08/24/2023).

Wei, Jason and Kai Zou (Aug. 2019). *EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks*. arXiv:1901.11196 [cs]. DOI: 10.48550/arXiv.1901.11196. URL: http://arxiv.org/abs/1901.11196 (visited on 08/24/2023).

Wild, Cameron T. et al. (Dec. 2019). “Media coverage of harm reduction, 2000–2016: A content analysis of tone, topics, and interventions in Canadian print news”. In: *Drug and Alcohol Dependence* 205, p. 107599. ISSN: 0376-8716. DOI: 10.1016/j.drugalcdep.2019.107599. URL: https://www.sciencedirect.com/science/article/pii/S037687161930376X (visited on 08/24/2023).

Zhang et al. (2021). “Social Media Exposure, Psychological Distress, Emotion Regulation, and Depression During the COVID-19 Outbreak in Community Samples in China”. In: *Frontiers in Psychiatry* 12. ISSN: 1664-0640. URL: https://www.frontiersin.org/articles/10.3389/fpsyt.2021.644899 (visited on 08/24/2023).

MSc Project - Reflective Essay

Project Title:	Optimisation of Good News Classification using Large Language Models
Student Name:	Hannah Melkemaryam Claus
Student Number:	220961895
Supervisor Name:	Dr Bruno Ordozgoiti
Programme of Study:	MSc Artificial Intelligence

Development of a framework for sentiment-aware news classification

Creating an entire framework that supports various model architectures and their connected processes was a challenging task. The entire endeavour can be split into sub-tasks as follows.

Choosing the right dataset

The first challenge was to find the right dataset to train and test the models on. For this project, a dataset with both news headlines and news articles was wanted. However, most openly accessible datasets only included either headlines or articles. However, the All the News dataset, that was used in the end, included all the necessary information. Originally, the dataset included 2.7 entries, but that was too much to process as a single person, hence, a smaller version of the same dataset was used with ca. 143,000 entries. Unfortunately, after analysing the dataset more closely, it became evident, that the period stretched over the entire American presidential candidacy, and election in 2016 and 2017. Although it was an event that was heavily discussed worldwide, as Donald Trump managed to split society into two opposing parties, the data showed that most of the words also appeared in a political context. Nevertheless, this was one of the only accessible datasets that delivered the desired data, hence, the risk of creating models that would not achieve the same accuracy in other countries than the US was accepted. Once the dataset was chosen, it became evident, that the model needed to be trained on either headlines or articles, not both together, as the needed computing units were not available. Hence, to achieve broader results, the headlines and articles were split into separate datasets, so the models could first be trained and tested on headlines and later on on articles.

Labelling the dataset

After applying one of the threshold-based sentiment analysis methods to label the data, the results showed ambiguous values. For instance, some headlines would use ambiguous words, which would make it harder for the thresholds to be accurately reached. Thus, to receive a higher accuracy with the labels, four different labelling methods were used, which were then combined into one aggregated label. However, as it is harder for these models to capture any deeper word similarities and contexts, the labels had to be constantly re-evaluated to ensure the correct training and testing of the classifiers.

Analysing word similarities

Due to the 8-page limit of the thesis paper, the inspected word similarities received through the Skip-gram model could not be added. However, they will be briefly discussed here:

- 'death': ['crutcher', 'madaya', 'charleston', 'earthquake', 'brutal', 'vicious', 'jumping', 'family', 'orleans', 'quake'],
- 'life': ['acknowledges', 'mold', 'gentrifying', 'searching', 'genital', 'dilemma', 'tory', 'primarily', 'dreamless', 'femme'],
- 'good': ['eternal', 'conservatism', 'way', 'biopic', 'hammer', 'rory', 'filter', 'fracturing', 'critique', 'lotto'],
- 'bad': ['typical', 'jonbenét', 'kanté', 'higgs', 'showering', 'refreeze', 'nader', 'bang', 'lb', 'joanne'],

- 'man': ['allegedly', 'jaffa', 'tulsa', 'hyena', 'skydiver', 'trash', 'bangalore', 'bystander', 'stepson', 'suspect'],
- 'woman': ['gym', 'microbeads', 'slims', 'bam', 'tit', 'punched', 'picasso', 'probe', 'oppresses', 'unconscious'],
- 'happy': ['cher', 'deadpool', 'mcqueen', 'wonderland', 'ghostbusters', 'tribute', 'shortsighted', 'lawrence', 'television', 'origin'],
- 'unhappy': ['reconsider', 'corzine', 'caving', 'smacking', 'unmitigated', 'disparate', 'wiping', 'secy', 'warmed', 'b—and'],
- 'obama': ['barack', 'shifted', 'policy', 'attend', 'grassley', 'rousey', 'administration', 'hiroshima', 'stayed', 'obamatrade'],
- 'trump': ['predictor', 'nationalism', 'homecoming', 'gasoline', 'mph', 'airstream', 'chameleon', 'remix', 'hospitalized', 'catchy'],
- 'book': ['siriusxm', 'album', 'retirement', 'mann', 'board', 'walton', 'oscarssowhite', 'faculty', 'co', 'adelson'],
- 'school': ['student', 'satanic', 'scapegoat', 'disability', 'yale', 'teacher', 'campus', 'quadruplet', 'twelve', 'tiffany'],
- 'sex': ['sexually', 'disorder', 'transplant', 'dna', 'abuse', 'kissing', 'inspire', 'poker', 'hyena', 'frugality'],
- 'apple': ['iphone', 'revenue', 'hindering', 'software', 'earbuds', 'qualcomm', 'bolster', 'oracle', 'workforce', 'wireless'],
- 'movie': ['batman', 'trailer', 'dyke', 'composer', 'creative', 'yuja', 'comic', 'song', 'superheroes', 'wbgo'],
- 'university': ['campus', 'professor', 'satanic', 'satanist', 'student', 'abuse', 'glorification', 'retweets', 'sprayed', 'mistakenly'],
- 'london': ['terror', 'swimwear', 'collude', 'kubrick', 'masood', 'vile', 'relentless', 'cleaver', 'transgender', 'heathrow'],
- 'russia': ['russian', 'ukraine', 'sanction', 'ceasefire', 'spy', 'bashar', 'treaty', 'picking', 'installed', 'missile'],
- 'army': ['civilian', 'ite', 'fighter', 'seal', 'navy', 'tribal', 'warplane', 'squadron', 'shi', 'lumber'],
- 'feminism': ['ghostbusters', 'thrilling', 'wright', 'winslet', 'rushdie', 'rihanna', 'segregationist', 'reinvigorated', 'poser', 'kung'],
- 'girl': ['boko', 'boy', 'schoolgirl', 'haram', 'nigerian', 'parking', 'chibok', 'cameroon', 'reunited', 'eco'],
- 'boy': ['mother', 'stripper', 'girl', 'pirate', 'sportswriter', 'tribeca', 'ankle', 'hopeful', 'gorilla', 'parent'],
- 'kim': ['jong', 'tsang', 'nam', 'kardashian', 'lived', 'toxin', 'sister', 'assassin', 'kooky', 'auction'],
- 'music': ['pop', 'ghostbusters', 'villa', 'photographer', 'corinne', 'opera', 'crawford', 'hijab', 'artist', 'towards'],
- 'woke': ['photobombs', 'chauffeur', 'lovely', 'unsellable', 'syracuse', 'included', 'winslet', 'dunst', 'smoky', 'ledecky'],
- 'attack': ['mosque', 'burkina', 'terrorist', 'victim', 'terror', 'aqap', 'akbar', 'faso', 'martyrdom', 'massacre'],
- 'terror': ['islamist', 'attacker', 'jihad', 'attack', 'flex', 'detains', 'terrorist', 'assailant', 'emirate', 'backlog'],
- 'christian': ['armenian', 'affiliate', 'cathedral', 'condemns', 'manbij', 'sharif', 'iftar', 'province', 'islamist', 'egypt'],
- 'muslim': ['islamist', 'mattress', 'hate', 'shiite', 'warmingâ€™hellish', 'condemns', 'cologne', 'ecuadorian', 'mogul', 'exhausted'],
- 'modern': ['infinite', 'marry', 'fuel', 'faux', 'klimt', 'playground', 'dicey', 'lyricist', 'crab', 'sky'],
- 'art': ['scoop', 'paulson', 'graffiti', 'pizza', 'delicious', 'learned', 'marvel', 'circle', 'foremost', 'xx'],
- 'election': ['presidential', 'macron', 'contracted', 'derek', 'butt', 'outcome', 'centrist', 'artan', 'played', 'latin'],
- 'bias': ['funding', 'execution', 'kirsten', 'religious', 'condom', 'viewed', 'waymo', 'product', 'fails', 'rehash'],
- 'ai': ['borrower', 'drag', 'penthouse', 'mimic', 'scribe', 'boating', 'amplify', 'iggy', 'arundhati', ‚stringent‘]}

The above list shows the ten most similar words used in the context of the main word. Word similarities reveal the semantic proximity between words. Words with high similarity scores are likely to share related meanings, indicating their association in language usage. A few controversial words can be seen above in context to relatively unambiguous words. The fact that the connected words regarding ‚terror‘, ‚attack‘, and muslim‘ are very similar, shows the systemic racism

underlying American news content. As already discussed in the paper, news shape the way we humans think and perceive the world. If the above words are the most similar, it means that people are more likely to automatically use these words in the same context. This shows the importance of deep data analysis and mitigation of bias.

Another example of bias is shown when looking at the word similarities for ‚university' and ‚school'. Both terms seem to have a connection to the word ‚satanic', which seems to hint at a viral incident combining these words in 2016/2017. During that period it might have been in the same context, but to common knowledge, ‚satanic' should not be in the top 10 most similar words of ‚university' or ‚school'.

Thus, it is very important to look at the dataset, mitigate the biases and re-evaluate what kind of news content is appropriate in our society.

Strengths and weaknesses

One of the prominent strengths of the paper is its extensive evaluation of various models. The research encompasses a wide array of techniques, ranging from different machine learning architectures and traditional classifiers to transformer-based models like DistilBERT. This comprehensive analysis not only offers a broad overview of model performance but also provides insights into their effectiveness across diverse methodologies. Additionally, the recognition of ethical and psychological considerations is emphasised. By addressing the ethical implications of sentiment analysis, particularly biases and societal impact, the research aligns itself with the ongoing discourse on the responsible deployment of machine learning in society. This dimension adds depth to the study, emphasising the importance of considering broader consequences beyond technical accuracy. However, the paper also presents some weaknesses that merit consideration. While biases are acknowledged, the exploration of bias mitigation techniques remains limited due to restricted access to more appropriate datasets. A more comprehensive investigation into strategies for mitigating biases could augment the research's contribution to ethical AI deployment. However, the above data analysis also highlights how important it is to include diverse voices in the decision-making processes so that racism and sexism can be eliminated.
While the paper alludes to future directions, a more explicit discussion of potential real-world applications and the implications of integrating sentiment analysis into decision-making systems could enhance its practical relevance. Moreover, a broader contextualisation of transformer-based models, such as DistilBERT, within the landscape of NLP could provide a more comprehensive understanding of their significance and limitations. It would be very interesting to see how this project can be further extended by creating a user interface. Platforms such as The Good News Network already provide ways to consume positive news, however, their selection process is manual and very subjective. Using this type of platform to create a personalised news provider could benefit people to choose for themselves what they define as positive or negative news. As can be seen, by the above word similarities, American news are very Western-oriented and favour white, Christian values that might not be appreciated by other countries and cultures.
Nevertheless, this project has its strengths and weaknesses that leave room for future improvements. With more time and more data, more complex models can be created that even incorporate international news or implement multi-lingual processes to include more languages.

Theory vs. Practice

As already discussed, the created models are trained and tested on a very biased and limited dataset. It will not have the same impact in theory as in practice. A dataset that incorporates the racism and sexism depicted in the above word similarities can have strong impacts on society, harming marginalised groups and leading to potential long-term consequences. However, for this thesis, the risks were accepted due to the limited access to more diverse datasets. Unfortunately, there is not a great variety of datasets to use. Additionally, due to this being a single person’s effort, the computing capabilities are not comparable to the resources of big companies that would have the power to implement these types of frameworks with more diverse data and more complex architectures. Thus, this was a valuable experience in terms of the performance of various machine learning solutions for text classification, however, if this concept were to be applied to

real-world settings, all lessons learned would be applied to create better and more appropriate results.

Awareness of legal issues and sustainability

Social and ethical issues have already been introduced above, however, there are a few legal considerations to be made. Sentiment analysis involves processing and analysing textual data, which often includes copyrighted material. Ensuring compliance with copyright laws and usage permissions is crucial to avoid legal ramifications. Moreover, when extending this project to include a user interface, it is important to stay transparent and let users choose how much data can be shared. Additionally, privacy laws dictate how user-generated content is collected, stored, and used. Adhering to data protection regulations safeguards user rights and prevents potential legal violations.

The sustainability of the developed models pertains to their longevity, relevance, and ecological impact. These models should remain applicable in the face of evolving language trends and societal dynamics. Continual updates and adaptations are essential to maintain model accuracy and contextual relevance over time. Moreover, considering the energy consumption associated with training and deploying large-scale models, optimising energy efficiency contributes to the sustainable development of these models. However, as mentioned before, the used dataset was heavily influenced by the American elections in 2016, thus, the data is now outdated and will not be able to properly represent word similarities after that period due to ever-changing events in society. To successfully implement these classifiers, they need to be constantly updated, by adding new datasets and repeating the training, testing, and optimising processes. As a result, the framework itself will have a longstanding application as it is independent and offers the tools to perform these updates.