Pertanika J. Sci. & Technol. 32 (4): 1609 - 1628 (2024)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

A Deep Learning-based Classification Model for Arabic News Tweets Using Bidirectional Long Short-Term Memory Networks

Chin-Teng Lin[1], Mohammed Thanoon[2] and Sami Karali[1,2]*

[1]*Australian AI Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, 2007, Australia*
[2]*Umm Al-Qura University, College of Engineering and Computing at Al-Lith, Makkah, 24382, Saudi Arabia*

ABSTRACT

This research develops a classification model for Arabic news tweets using Bidirectional Long Short-Term Memory networks (BiLSTM). Tweets about Arabic news were gathered between August 2016 and August 2020 and divided into five categories. Custom Python scripts, Twitter API and the GetOldTweets3 Python library were used to collect the data. BiLSTM was used to train and test the model. The results indicated an average accuracy, precision, recall, and f1-score of 0.88, 0.92, 0.88, and 0.89, respectively. The results could have practical implications for Arabic machine learning and NLP tasks in research and practice.

Keywords: Arabic, Arabic dataset, Arabic news, ML, NLP, Twitter

INTRODUCTION

Social media platforms have changed the way people talk to each other and are now an important part of everyday life. Panagiotou et al. (2016) asserted that with the rise of social media, events can now reach a wider audience and create a buzz that was not possible earlier. Twitter users' tweets are a large source of unorganized and varied information. Aslam (2018) and Guzman et al. (2017) stated that around 326 million individuals log onto Twitter monthly and generate around 500 million tweets. That is the equivalent of 6000 tweets per second. Tweets were initially intended as more of a light-hearted medium, but as researchers saw their potential, they quickly became one of the most researched platforms. Twitter does not enforce rules for its users, so anybody can post whatever they want.

ARTICLE INFO

Article history:
Received: 16 July 2023
Accepted: 01 February 2024
Published: 16 July 2024

DOI: https://doi.org/10.47836/pjst.32.4.09

E-mail addresses:
mithanoon@uqu.edu.sa (Mohammed Thanoon)
chin-teng.lin@uts.edu.au (Chin-Teng Lin)
sami.karali@student.uts.edu.au (Sami Karali)
* Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

Tweets contain a wealth of data, yet some could be incorrect. In post-Trump politics, false news has taken the central stage, and social media platforms like Facebook and Twitter have been linked to its proliferation (Hunt, 2016). Automatically categorizing tweets is necessary to improve information retrieval for individual and institutional needs (e.g., news filtering, false news tracking, rumor mining). That is why training machine learning algorithms to properly categorize tweets is more important than ever. These algorithms may spot and remove hoaxes, rumors, and other nonsense from users' Twitter feeds.

Many challenges come with classifying Arabic tweets because of their vast potential. Another challenge is that Twitter has limited words for tweets, sometimes leading to ambiguous statements and unclear expressions, which can result in misclassifications in practical situations. Moreover, the Arabic language is multifaceted, including regional dialects and colloquialisms, adding another layer of complexity.

Natural Language Processing (NLP) methods like sentiment analysis, news categorization, and identifying rumor news can also analyze tweets. Alonso et al. (2021) observed that sentiment analysis is a good way to find and fight fake news. Kowsari et al. (2019) opined that putting news tweets into categories based on what they are about could help improve the quality of information people get and help the media share what news people are most interested in. People who work in the news will look at tweets and people's interests to get a picture of the population. When researchers classify texts, they can easily look for relevant information, look at it, and organize it for future use. Manual text categorization is possible, but it is a laborious process that takes a lot of effort and time. Text classification, which uses machine learning and deep learning, is a great way to automatically identify or classify text like news.

Many studies used machine-learning techniques (Ikonomakis et al. (2005); Buabin (2012). Yang and Pederson (1997) stated that classification is the process of putting a piece of text into one of many predetermined groups based on the information it contains. Al Sbou et al. (2018) stated that, as 6.6% of the world's population speaks Arabic, it is one of the most widely used languages. It is the fifth most popular language on the web, according to Albalooshi et al. (2011) and comes in three different forms, one of which is classical Arabic, which is used in religious and ancient scripts. A report by Raftery (2017) reveals that two-thirds of young Arabs use social media platforms like Facebook and Twitter as their primary sources of news, and the area is home to eleven million monthly active Twitter users who send out over 27.4 million tweets daily. Arabic tends to attach prepositions, pronouns, and the article to words. It makes the language very inflectional and agglutinative. It is important to know how they are put together syntactically to represent and change Arabic words to be used in a categorization system.

Arabic text categorization has received comparatively less attention than its English counterpart. Very few studies have focused on the categorization of short texts in Arabic.

In addition to the linguistic peculiarities of Arabic, the lack of open access to a small text corpus in the language is a limitation (Al-Tahrawi & Al-Khatib, 2015). There are only four other known papers that attempt to categorize Arabic tweets. Deep learning was utilized by Bdeir and Ibrahim (2020), and classical machine learning was employed by Bekkali and Lachkar (2014), Abdelaal et al. (2018) and Ibrahim et al. (2021).

As more and more Arabic news tweets are shared daily, the question now arises: How can we effectively classify and analyze them to meet the increasing need for accurate and reliable NLP models? By developing a deep learning-based categorization model for Arabic news tweets with Bidirectional Long Short-Term Memory (BiLSTM) networks, this research presents an approach to overcoming this challenge. By collecting and organizing tweets into broad categories like "general news," "regional news," "sports news," "economic news," and "quality of life news," we were able to train and test our model to adequate levels across a variety of metrics, including accuracy, precision, recall, and f1-score.

Our research can aid in producing more targeted and relevant news material for Arabic speakers, and it can provide significant assistance for the development of natural language processing systems by providing insights into the precise categories of information that the population seems to be most interested in.

RELATED WORK

According to Salloum et al. (2017a, 2017b), Alabbas et al. (2016), Salloum et al. (2018), and Elhassan and Ahmed (2015), automating the process of categorizing a collection of documents according to their content using technologies and algorithms is a text classification process. Elhassan and Ahmed (2015) noted that it is a technique for locating and navigating big datasets and organizing them into meaningful categories for later use. There has been a rise in specialized studies on the classification of texts as massive data are available from numerous sources, including websites, emails, news stories, social media posts, reports, and journals.

Twitter Dataset for NLP

This study aims to learn how and why tweeted datasets have become so widely used for studying people's opinions and responses. Twitter data is more useful than Facebook data for building 50 large corpora for natural language processing, as was pointed out by Ahmed et al. (2017). However, there is a dearth of data compared to the English dataset, as observed by Assiri et al. (2018). Existing dialects of Arabic, which vary from modern standard Arabic and even between Arab countries, further complicate matters. Data mining for Arabic sentiment analysis has greatly benefited from the efforts of Almuqren and Cristea (2021), which have created the gold-standard Saudi corpus AraCust and the Saudi lexicon AraTweet (ASA). AraSenCorpus is another repository of information; it includes 4.5 million tweets in both

standardized and colloquial varieties of Arabic. The widespread relevance and practicality of this dataset demonstrate the value of data. When organizing human opinion into a data set, Arabic Sentiment Analysis (ASA) and Arabic Text Classification (ATC) are crucial tools. Sentiment analysis is a method that can be used to calculate how the public feels about a topic. Feng and Kirkley (2021) researched the emotional response of the population to policies developed for dealing with COVID-19, and their findings suggest that tweets can be used to improve public health monitoring and crisis management. Jordan et al. (2018) determined that utilizing Twitter as a data source yields real-time information from various geographical regions. However, there is a lack of investigation into the potential of Twitter data for tracking and anticipating the public's reactions to issues. More research is required to validate the utility of datasets across disciplines.

Deep Learning Approaches

Encoding textual features using bag-of-words, Ngrams, or Tf-IDF is common in text classification tasks, including Arabic. On the other hand, recent neural network-based models have dramatically surpassed these older approaches. Standard approaches, including gated neural network models like LSTM by Schmidhuber and Hochreiter (1997) and GRU by Chung et al. (2014), were able to circumvent these issues largely thanks to deep learning techniques. To consider context, Vaswani et al. (2017) observed that the Transformer Network architecture with self-attention layers has replaced all recurrent and convolutional layers as the state-of-the-art in many NLP tasks. It inspired the development of novel neural network architectures for contextual word embedding models like ELMO by Peters et al. (2018), ULmFit by Howard and Ruder (2018), and new hybrid Bidirectional Encoder Representations from Transformers (BERT) architectures by Devlin (2018). Even though deep learning models are powerful, many challenges may affect the model's performance, especially related to data quality on platforms like Twitter, due to brevity and ambiguity. Researchers have been addressing the scalability of these models in terms of both computational efficiency and adaptability to evolving news topics.

Self-attention layers shrank the model, improved training efficiency, and produced outstanding results in neural machine translation. Deep learning methods have been used in many areas of study, such as Arabic NLP and sentiment analysis. El-Alami and Alaoui (2016) used deep learning to improve Arabic text classification. They used a deep stacked auto-encoder and short reproduced codes to reduce the number of dimensions in the representation space. Sayed et al. (2017) examined how well Arabic text can be sorted using a deep neural network based on textual similarity and N-gram level. The results showed a deep learning classifier with an AR of 98.50%, a 75% similarity level, and a 3-gram outperforms the SVM, NB, and k-NN classifiers. Using a deep learning strategy, Boukil et al. (2018) improved the performance of their Arabic text classifier. They found 111,728

news documents on the web and put them into five categories to make a corpus. They used the TF-IDF method to choose which features to use and a vector-words strategy to show the text. With an accuracy of 92.94%, the CNN model outperformed both LR and SVM.

Galal et al. (2019) used CNN (Convolutional Neural Network) deep learning to classify Arabic texts. As a preprocessing step, they sorted all the Arabic words with the same root into groups using the proposed Gstem technique. The accuracy of the CNN improved to 92.42% with the help of Gstem from 88.75% without it. Elnagar et al. (2020) proposed and evaluated nine deep-learning methods for classifying single- and multi-label Arabic text. SANAD and NADiA, two massive Arabic text collections, were used to validate Arabic text classifiers. The best models for single-label classification tasks were HANGRU, CNN, BiGRU, and BiLSTM. These models have since been trained for multi-label classification. The models were evaluated based on their accuracy using the accuracy metric.

Arabic Text Classification

Because of the nature of the Arabic language and the scarcity of adequate resources, categorizing Arabic text is more difficult. There is a big difference between the verbal and nominal sentences, where the latter do not need a verb, and the vocabulary size, and between the use of diphthongs and long vowels. A significant factor in using a particular methodological framework is the nature of classification, which may involve short versus long documents, multiclass versus binary classification, or multi-label classification. Studies of text classification in Arabic are in tandem with suitable benchmarks and corpora development.

Researchers have suggested many methods and approaches for classifying Arabic text. Using support vector machines, Moh'd Mesleh (2011) investigated the value of feature sub-set selection metrics and provided an empirical comparison of them (SVM). Hmeidi et al. (2015) studied the performance of such classifiers and examined how using different Arabic stems, like the light and root-based ones, affected the results. Embedding-based methods, such as word average and document embeddings like Doc2-vec and Glove, have been proposed by El Mahdaouy et al. (2016). Convolutional and recurrent neural networks have seen a rise in popularity and research due to their ability to model sentences based on sequences of context windows and capture local correlations. In contrast to RNN's ability to treat sequences of any length and capture long-term dependencies, CNN's emphasis on features at different sentence positions through convolutional filters and pooling sets it apart.

Dahou et al. (2016) proposed building a CNN for Arabic sentiment analysis on top of a model that used web-crawled corpus words to train embeddings for Arabic words. Both Alayba et al. (2018) and El-Alami et al. (2020) investigated convolutional neural networks (CNNs) and long short-term memories (LSTMs) and their hybrids for classifying Arabic text using a retrofitting technique that makes use of the semantic information embedded in Arabic Word-Net. Contextual embedding models like ELMo, ULMFiT, and BERT are

current developments in language modeling. The pre-trained Ara-BERT for Arabic was introduced by Antoun et al. (2020), and it was tested on a variety of natural-language understanding tasks and compared to the multilingual BERT. A large, evenly distributed Arabic short-text dataset is still lacking. Classifying news tweets is a challenging task related to short text classification tasks (Mohammed et al., 2020). Numerous languages, including Arabic, exhibit this property, as observed by Khoja et al. (2017). Despite recent progress in Arabic text classification, challenges like precise semantic capturing arise due to the complexity and ambiguity of Arabic text data.

METHODOLOGY

A model was created with real data for news tweet topic classification. A hierarchical approach has been built and divided into four phases, as shown in Figure 1.

Tweets Harvesting Phase

The first phase is news tweet collection. All news tweets were collected using self-made Python scripts for both the Twitter application programming interface (API) and the GetOldTweets3 Python library, allowing historical tweets to be collected beyond the 1-week limitation of the standard Twitter API. A larger set of tweets with a more diverse set enhances the robustness of the model.

Preprocessing Phase

The second phase is cleaning the tweets of unwanted elements like non-Arabic words, symbols, URLs, emoticons, and stop words, especially those common in Arabic. The rationale of this phase is data cleaning to ensure that the model is trained on relevant content, improving its accuracy and efficiency. After collecting the news tweets, the preprocessing process was conducted. Only root tweets were considered, and retweets were excluded. All news tweets other than those in Arabic were excluded, including numbers. Then, all punctuation marks were removed. Furthermore, any URLs and emoticons were removed to obtain higher textual quality. Stop words were also excluded.

Features Extraction Phase

The feature extraction phase is the process of transforming raw data into numerical features. This phase is responsible for a linear combination of the existing features. The recent state-of-the-art model, the AraBERT model, was used in this work. This model is a pre-trained BERT that is specifically for the Arabic language. This model tokenizes sentences and converts each token into embedding vectors. AraBERT is designed specifically for Arabic, which captures linguistic nuances that generic models may miss.

Classification Phase

The final phase is decision-making, also known as the classification phase. The classifier in this phase acts on the sentence representation to get the decision and give each tweet a label belonging to one of the predefined classes. Various techniques can be used for text classification. Graves and Schmidhuber (2005) found BiLSTM suitable for text classification due to its effectiveness in NLP as well as its use of memory as specific hidden units. The BiLSTM model was implemented with specific configurations, including an embedding size of 256 words, a training batch size 64, and 5 epochs. BiLSTM is effective in text classification because of its ability to remember long-term dependencies, which makes it suitable for tweet classification.

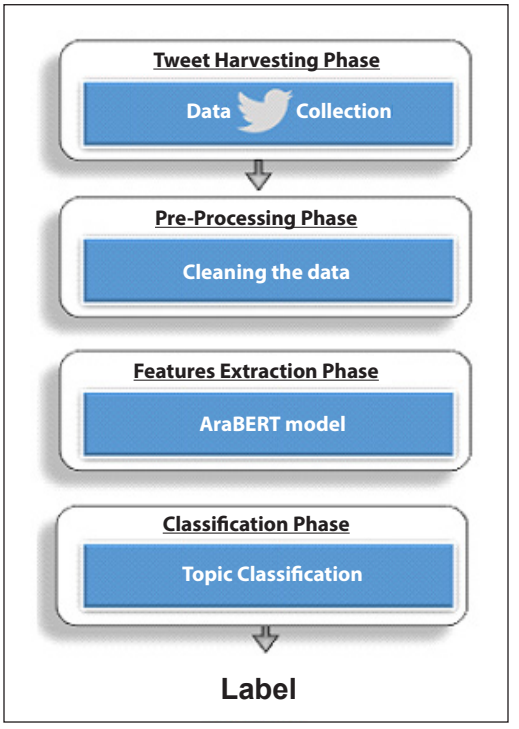

Figure 1. The flowchart of the model

DESIGN AND IMPLEMENTATION

Tweets Harvesting Phase

The target was to build a large Arabic news tweet dataset for Saudi Arabia. The dataset contained 89,179 news tweets related to Saudi Arabian news. News tweets were collected from verified users on Twitter, and all users were official news agencies of Saudi Arabia. These users had already categorized themselves into five classes: General News, Regional News, Sports News, Economic News and Quality of Life News. Moreover, they all posted tweets based on the class they belonged to. Therefore, each news tweet was labeled based on the user class from which the tweet was collected.

All news tweets were collected using self-made Python scripts for the Twitter application programming interface (API) and the GetOldTweets3 Python library. The Twitter API is a back-end server that stores all individuals' tweets. API provides a service where data can be collected for the public, but it has the limitation that data more than 1-week old cannot be accessed, as revealed by Search Tweets (https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/guides/standardoperators). However, there are many methods to extract the data that can be used, and there are no restrictions for collecting many tweets and accessing the historical tweets, like the GetOldTweets3 Python library

used in this study. GetOldTweets3 is an open-sourced Python library that can be used with specific keywords and periods to extract historical tweets, as averred by Jefferson (2018).

The collection period ranges from the day of the user creation (in general, August 2016) to August 1, 2020. Data were collected at different periods for each of the labels. Arabic filter was used to collect all data. Only root tweets were considered, and retweets were excluded. The data were formatted as Excel text files.

Preprocessing Phase

This phase is one of the most important processes that go into a given dataset to obtain a clean version and prepare the dataset for the next phase. Even though the preprocessing phase is crucial for enhancing data quality, it might result in the loss of some relevant information. The dataset used in this research is substantial, reflecting the wide range of news topics. However, this diversity and the ever-evolving nature of news suggest that the dataset might still lack certain nuances.

After collecting the news tweets, the preprocessing process was conducted. Only root tweets were considered, and retweets were excluded. All news tweets other than those in Arabic were excluded, including numbers. Then, all punctuation marks were removed. Furthermore, any URLs and emoticons were removed to obtain higher textual quality. Stop words were also excluded.

The collected news tweets are publicly available (Karali et al., 2021). They were formatted into an Excel file containing labels, and all tweets were listed under their label with the posted date. The dataset contained only tweet IDs with their labels and posted days without any personally identifying information, using the user IDs to comply with Twitter's Developer Policy for the dataset, as extracted from Twitter's Developer Agreement and policy (https://developer.twitter.com/en/developer-terms/agreement-andpolicy). Table 1 presents one example of the raw data before and after preprocessing and its translation.

Table 1
One example of the raw data and its translation is before and after preprocessing

Before processing	
Tweet	Translation
عاجل #مجلس_الوزراء: الموافقة على اتفاقية# الخدمات الجوية بين حكومة المملكة وحكومة آيسلندا.	# Urgent # Council of Ministers: The Air Services Agreement between the Kingdom of Saudi Arabia and the Icelandic governments has been approved.
After processing	
Tweet	Translation
عاجل مجلس وزراء موافق اتفاقي خدم جوي حكوم مملك حكوم آيسلندا	Urgent Council Ministers Air Service agreement Kingdom Saudi Arabia government Icelandic government approved.

Processing involved the removal of hashtags, prepositions, definite and indefinite articles, conjunctions and participles. The change in the content due to such processing is evident in Table 1.

Features Extraction Phase

The feature extraction phase is the process of transforming raw data into numerical features. This transformation aims to process the numerical features and preserve the original dataset. The recent state-of-the-art model, the AraBERT model, was used in this work. The AraBERT model was used because it has proved its effectiveness with deep learning models, such as the BiLSTM model. This model is a pre-trained BERT that is specifically for the Arabic language. AraBERT uses the attention mechanism (Antoun et al., 2020). This mechanism learns contextual relations between words. Table 2 shows the structure of all versions of AraBERT **(**Matrane et al., 2021), and AraBERT base V2 was used in this study. AraBERT breaks up the sentences into sections (tokenization), and this method is relevant for Arabic as it is based on the Farasa Segmenter, as illustrated by Saeed (2021). After that, each token was converted to weights and embedding vectors. Table 2 summarizes the architectures of the AraBERT model **(**Elfaik & Nfaoui, 2021).

Table 2
AraBERT architectures details (Elfaik & Nfaoui, 2021)

	AraBert	ArabicBERT Mini	ArabuBERT Medium	ArabicBERT Base	ArabicBERT Large
Hidden layers	12	4	8	12	24
Attention heads	12	4	8	12	16
Hidden size	768	256	512	768	1024
Parameters	110 M	11 M	42 M	110 M	340 M

ArabicBert Large can identify 24 layers,16 attention heads and a hidden size 1024 using 340 parameters. Thus, it is the most efficient feature extraction tool.

In Figure 2, a news tweet is divided into two parts: (1) sentence embedding and (2) positional embedding. AraBERT uses an encoder for each part of the news tweet and converts it into a vector. Then, in an activation layer, AraBERT uses a transformer decoder to calculate the score of each part.

Classification Phase

Various techniques can be used for text classification. Graves and Schmidhuber (2005) found BiLSTM suitable for text classification due to its effectiveness in NLP as well as its use of memory as specific hidden units. In text classification, storing context history information and recalling the input words will support the model for classification

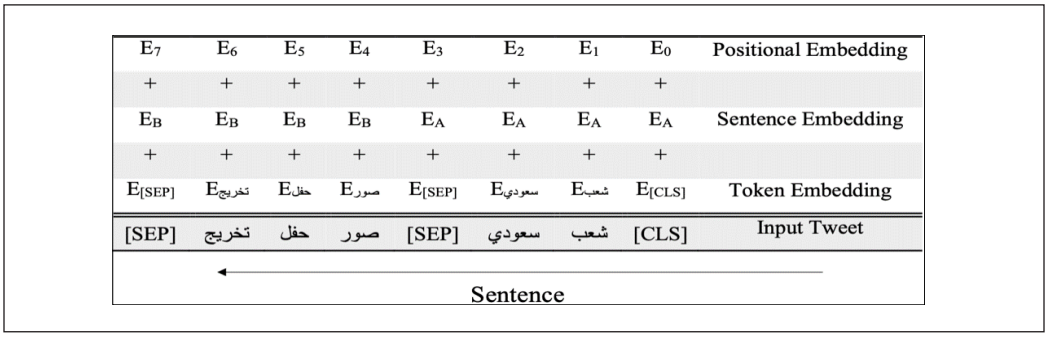

Figure 2. AraBERT model example of creating the input: Token, position, and sentence embeddings

and get higher results. BiLSTM has these characteristics, so this classifier was used in this research.

The BiLSTM model is based on the sequence-to-target concept and bidirectional architecture. BiLSTM makes the sequence information of a neural network in both directions, from front to back and in reverse. In regular LSTM, the input flow is in one direction, which is either forward or backward. On the other hand, both directions for the input flow in BiLSTM preserve past and future information.

Table 3
BiLSTM layer configuration and output shapes

Layer (type)	Output Shape
embedding (Embedding)	(None, 10, 256)
dropout (Dropout)	(None, 10, 256)
Bidirectional (Bidirectional)	(None, 10, 512)
dropout_1 (Dropout)	(None, 10, 512)
bidirectional_1 (Bidirectional)	(None, 256)
dropout_2 (Dropout)	(None, 256)
flatten (Flatten)	(None, 256)
dense (Dense)	(None, 32)
dropout_3 (Dropout)	(None, 32)
dense_1 (Dense)	(None, 5)

Figure 2 illustrates the structure of the model. The embedding feeds the BiLSTM layer, and the dropout layer follows it to avoid overfitting. After that, dense layers are the following layers. The BiLSTM model is built using the following characteristics: 256-word embedding size, 64 training batch size, five epochs, 10 Max sentence length and Adam optimizer. In this study, after the BiLSTM model was built, it was applied to the dataset that has been collected to classify each news tweet to the class to which it belongs. Table 3 shows the BiLSTM text classification used in this research. The BiLSTM model started with an embedding technique to produce 256-dimensional word vectors for each word.

The pseudocode, presented later, provides a clear and structured representation of the methods used in this research.

RESULTS

Machine learning models can be evaluated in several ways to see how well they perform. Automation-based evaluation is one such method. It employs objective metrics like recall, precision, and f1-score to conclude the quality of an evaluation's results. In this research,

we used this method to test how well our algorithm could divide tweets into five distinct categories: (1) General News, (2) Regional News, (3) Sports News, (4) Economic News, and (5) Quality of Life News. For this purpose, we split our dataset into three parts: (1) training set, (2) validation set, and (3) testing set. The main goal was to train the model to properly categorize each tweet by its related label. After applying the BiLSTM model to the dataset, we used several methods to assess our classification model's efficacy. These included a receiver operating characteristic (ROC) curve, a confusion matrix, and a graphical representation of the false positive rate (FPR) at varying thresholds. Table 4 shows the calculation of the class-wise metrics for each class and the overall accuracy of 88%.

Table 4
Calculation of the class-wise metrics for each class and the overall accuracy

Class	Precision	Recall	F1 score	Support
Economic News (C0)	0.99	0.97	0.98	688
General News (C1)	0.88	0.97	0.92	19276
Quality Life News (C2)	0.98	0.98	0.98	293
Regional News (C3)	0.79	0.47	0.59	4937
Sports News (C4)	0.97	0.99	0.98	1560
Accuracy		0.88		

Among the five categories, all evaluation variables were good for economic news, quality of news and sports news, as the values for precision, recall and f1-score were in the range of 097 to 0.99.

Furthermore, the area under the ROC curve (AUC) of 0:86 in Figure 3 demonstrates the model's overall performance, showing that it can distinguish between the positive and negative classes. The model's accuracy was broken down into true positives and false negatives predicted samples, as shown in the confusion matrix in Table 5. Additional insight into the model's performance was offered by the graphical representation of the FPR at different classification criteria (Figure 3), which showed that the model maintained a relatively low FPR even at high classification thresholds.

Table 5
Normalized confusion matrix for a multiclass dataset

	Confusion Matrix					
	C_0	C_1	C_2	C_3	C_4	1.0
C_0	0.98	0.02	0	0.02	0	0.8
C_1	0.01	0.97	0.01	0.04	0.01	0.6
C_2	0	0.02	0.98	0	0.01	0.4
C_3	0.01	0.53	0.01	0.48	0.01	0.2
C_4	0	0.02	0	0	0.99	0.0

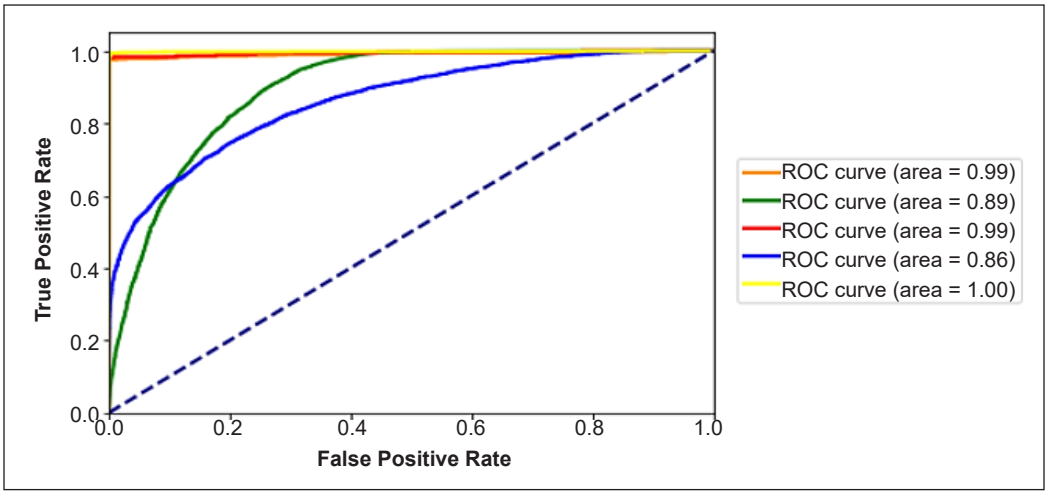

Figure 3. Receiver operating characteristic (ROC) (AUC) curves fitted to the data

Our results indicate that the model successfully classified the samples into their intended categories. Although there were a few false positives, the overall performance was good, as measured by a high AUC and a low FPR. The results demonstrate that the model performs reasonably well in accurately distinguishing between the positive and negative samples. However, there is room for improvement, particularly in reducing the number of false positives and false negatives.

ANALYSIS OF ERRORS AND LIMITATIONS

While our model demonstrates promising results in classifying Arabic news tweets from Saudi Arabia, it is important to acknowledge its limitations and potential sources of errors:

1. Sources of False Positives and Negatives: False positives can arise from the inherent ambiguity of some tweets, where content may overlap between categories. For instance, a tweet discussing the economic impact of a regional sporting event might be misclassified between "Economic News" and "Sports News." On the other hand, false negatives might result from unique linguistic structures or vernacular language for which the model has not been adequately trained.
2. Data Quality and Quantity: Data quality plays a pivotal role in the performance of any machine learning model. Although we sourced tweets from official, accredited news agencies, biases or inaccuracies in the data are always possible. Additionally, the volume of data, especially for underrepresented categories, might not be sufficient to capture all linguistic nuances.
3. Generalizability and Scalability: Our model is designed to fit Arabic news tweets from Saudi Arabia. This specificity ensures higher accuracy for this dataset, but it might not generalize well to tweets from other Arabic-speaking regions with different dialects

and sociocultural references. As for scalability, while the model efficiently handles the current dataset, its performance with much larger datasets remains untested.

4. Ethical and Social Implications: In Saudi Arabia's unique cultural, social, and political landscape, the ethical and social implications of a model analyzing tweets are crucial. For instance, consider a scenario where the model classifies tweets discussing an important cultural event in Riyadh as "general news" instead of "regional news." Furthermore, even though tweets are public, users in Saudi Arabia might have distinct privacy expectations. It is vital to ensure data collection respects these cultural norms and values. There is also potential regional bias if the model predominantly learns from tweets from specific Saudi regions, possibly leading to a skewed representation of national sentiments.
5. Model Limitations: While BiLSTM is effective, it has limitations. It might not capture very long-term dependencies in text as efficiently as some other architectures. Moreover, the model's complexity can lead to extended training times, especially with larger datasets.

Addressing these limitations and potential sources of errors is crucial for the model's applications and future iterations. In this research, we aim to pave the way for future models that are more resilient and inclusive.

DISCUSSION

The research developed the model to classify Arabic news tweets in Saudi Arabia. This primary focus brings forth certain considerations. The linguistic and cultural distinctions of Saudi Arabia are different in many ways, even though Saudi Arabia represents a significant portion of the Arabic-speaking world. Due to this specificity, the model's performance may be influenced when exposed to tweets from other Arabic-speaking regions, as each region has its unique dialects and sociocultural references. Therefore, it is important to consider how well the model in this research would generalize and adapt to such diverse content. Moreover, tweets from verified and official news agencies that disseminate news in this research carry ethical weight.

News agencies on platforms like Twitter provide content intended for public consumption. The content often includes official statements, reports, or sensitive topics. Leveraging this kind of data for research purposes requires careful consideration. Ensuring the context, nuances, and intent behind the tweets are accurate is important. Misinformation or bias, even if unintentional, can have significant implications given the official nature of these sources. Therefore, researchers are responsible for approaching this data with the utmost integrity.

After identifying broader considerations, it is crucial to delve deeper into specific performance metrics and insights derived from our model. The results highlighted the

model's effectiveness in classifying Arabic news tweets from Saudi Arabia and pointed out areas for improvement, taking into account the challenges and ethical concerns mentioned above.

The classification of news articles is essential for machine learning and natural language processing tasks. Analyzing user-generated content, such as tweets, has become more crucial as social media platforms have proliferated. This paper used bidirectional long short-term memory (BiLSTM) networks to develop a deep learning-based classification model for Arabic news tweets. Analysis of the collected data suggests that we go over the categories of news stories that the Arabic-speaking population is most interested in, as evidenced by Rey (2019).

We investigated various definitions of general, regional, sports, economic, and quality of life news and how they relate to the Saudi people's interests and priorities. We also identified how our findings could be used in real-world applications like machine learning modeling and natural language processing.

We compiled a dataset of tweets about Arabic news from August 2016 to August 2020 and categorized them into five groups to achieve this goal. Custom Python scripts were used with the Twitter API and the GetOldTweets3 Python library to collect these data. According to our analysis, each news category has a significant relationship with Saudi Arabia's Vision 2030. In what follows, we will summarize the results of our study that looked at the news that Saudi Arabians tweeted about. The study examined the population's news preferences, revealing its interests. Now, let us put each paragraph into a category based on the type of news it contains:

General News

According to the results of our study, the general news is what the population of Saudi Arabians is most interested in. Although Lehman-Wilzig and Seletzky's (2010) definition of general news as an intermediate category of news between hard (political, economic, or social topics) and soft (gossip, local scandal, and human interest), Rey (2019) defines general news as the most significant local and international news typically found on the front page with a big, bolded title called «banner headline,» and this is more appropriate for our study. We discovered that the Arabic population reads this news category the most.

Regional News

This study distinguishes local/regional news as a distinct sector for analysis, even though national/international news does cover some local urgent or top events. «regional news» refers to feature articles that draw attention to relevant local events—the discussion centers on factors that may influence the population, including their choices and actions.

Economic News

For many reasons, Saudis are concerned about the economy. The state's economy is extremely reliant on oil and its sales in the global market, as per the IMF (2016) report on economic diversification in oil-exporting Arab countries presented in the *Annual Meeting of Arab Ministers of Finance, Manama, Bahrain*. Washington, DC: IMF. The recent fluctuations in oil prices and consumption may worry the locals because they know the importance of oil. The fact that so many tweets deal with economic topics demonstrates that citizens of Saudi Arabia are interested in and eager to learn about the state of the economy and any developments in this regard.

Quality of Life News

One of the goals of the Saudi Arabian government, Vision 2030, is to improve the citizens' standard of living. Changes in ecosystems, cultures, environments, and sports are highlighted as priorities in Saudi Vision 2030 (https://www.vision2030.gov.sa). The government's efforts to entice its citizens into these areas are motivated by a desire to raise the standard of living for its citizens. Participation in state life in various fields guarantees increased opportunities for people. According to a Ministry of Foreign Affairs report, social, cultural, and other services that improve quality of life are guaranteed. Therefore, it is reasonable and commendable that people care about hearing about improvements to their quality of life.

Sports News

News about sports, the economy and quality of life have all been criticized for being uninteresting to people's needs. The people of Saudi Arabia care more about sports than they do about improving their standard of living or their country's economy. It might be explained by the government's heightened interest in sports development and expanding female sports nationwide, one of Vision 2030's objectives. Even though only a small portion of Saudi adults engage in physical activity, the local population actively participates in sports as spectators at the national and international levels.

CONCLUSION AND FUTURE WORK

In conclusion, based on an analysis of what people in Saudi Arabia tweeted about, this study has given us important information about what kind of news they like. People in Saudi Arabia are most interested in reading about the country's progress toward the general, regional, sports, economic, and quality of life news goals set out in Saudi Vision 2030. It is shown by data collected from August 2016 to August 2020 using custom Python scripts and the Twitter API.

The dataset from this study can be used for many natural language processing (NLP) tasks and machine learning models. It makes it useful for the scientific community, educational institutions, students, and researchers who need a record of formal news from a certain time. Additionally, Bidirectional Long Short-Term Memory (BiLSTM) networks have been used to classify the dataset, leading to a model with a macro-average precision of 0.92, recall of 0.88 and f1-score of 0.89.

In terms of future work, more research could look at the news preferences of the Saudi Arabian population by looking at how they use social media and news sources other than those used in this study. Further understanding of the news' function in Saudi society may be gleaned by examining the influence of news sentiment on the public's attitudes and actions.

ACKNOWLEDGEMENT

The authors sincerely thank University of Technology Sydney, Australia, for providing an enriching and supportive environment that greatly facilitated my academic endeavors and research.

REFERENCES

Abdelaal, H. M., Elmahdy, A. N., Halawa, A. A., & Youness, H. A. (2018). Improve the automatic classification accuracy for Arabic tweets using ensemble methods. *Journal of Electrical Systems and Information Technology*, *5*(3), 363-370. https://doi.org/10.1016/j.jesit.2018.03.001

Ahmed, W., Bath, P. A., & Demartini, G. (2017). Using Twitter as a data source: An overview of ethical, legal, and methodological challenges. In K. Woodfield (Ed.), *The Ethics of Online Research (Advances in Research Ethics and Integrity,* (Vol. 2, pp. 79-107). Emerald Publishing Limited. https://doi.org/10.1108/S2398-601820180000002004

Al Sbou, A. M., Hussein, A., Talal, B., & Rashid, R. A. (2018). A survey of Arabic text classification models. *International Journal of Electrical and Computer Engineering, 8*(6), 4352-4355. https://dx.doi.org/10.11591/ijece.v8i6.pp4352-4355

Alabbas, W., Al-Khateeb, H. M., & Mansour, A. (2016). Arabic text classification methods: Systematic literature review of primary studies. In *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt)* (pp. 361-367). IEEE Publishing. https://doi.org/10.1109/CIST.2016.7805072

Alayba, A. M., Palade, V., England, M., & Iqbal, R. (2018). A combined CNN and LSTM model for Arabic sentiment analysis. In A. Holzinger, P. Kieseberg, A. Tjoa, E. Weippl (Eds.), *Machine Learning and Knowledge ExtractionCD-MAKE 2018, Lecture Notes in Computer Science* (Vol 11015, pp. 179-191). Springer. https://doi.org/10.1007/978-3-319-99740-7_12

Albalooshi, N., Mohamed, N., & Al-Jaroodi, J. (2011). The challenges of Arabic language use on the Internet. In *2011 International Conference for Internet Technology and Secured Transactions* (pp. 378-382). IEEE Publishing.

Almuqren, L., & Cristea, A. (2021). AraCust: A Saudi Telecom tweets corpus for sentiment analysis. *PeerJ Computer Science*, *7*, Article e510. https://doi.org/10.7717/peerj-cs.510

Alonso, M. A., Vilares, D., Gómez-Rodríguez, C., & Vilares, J. (2021). Sentiment analysis for fake news detection. *Electronics*, *10*(11), Article 1348. https://doi.org/10.3390/electronics10111348

Al-Tahrawi, M. M., & Al-Khatib, S. N. (2015). Arabic text classification using Polynomial Networks. *Journal of King Saud University-Computer and Information Sciences*, *27*(4), 437-449. https://doi.org/10.1016/j.jksuci.2015.02.003

Antoun, W., Baly, F., & Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. *ArXiv,* Article 2003.00104. https://doi.org/10.48550/arXiv.2003.00104

Aslam, S. (2018). *Twitter by the numbers: Stats, demographics & fun facts*. Omnicoreagency. *com*. https://www.omnicoreagency.com/twitter-statistics/

Assiri, A., Emam, A., & Al-Dossari, H. (2018). Towards enhancement of a lexicon-based approach for Saudi dialect sentiment analysis. *Journal of Information Science*, *44*(2), 184-202. https://doi.org/10.1177/0165551516688143

Bdeir, A. M., & Ibrahim, F. (2020). A framework for Arabic tweets multi-label classification using word embedding and neural networks algorithms. In *Proceedings of the 2020 2nd International Conference on Big Data Engineering* (pp. 105-112). ACM Publishing. https://doi.org/10.1145/3404512.3404526

Bekkali, M., & Lachkar, A. (2014). Arabic tweets categorization based on rough set theory. *International Journal of Computer Science & Information Technology*, *6*, 83-96. https://dx.doi.org/ 10.5121/csit.2014.41109

Boukil, S., Biniz, M., El Adnani, F., Cherrat, L., & El Moutaouakkil, A. E. (2018). Arabic text classification using deep learning technics. *International Journal of Grid and Distributed Computing*, *11*(9), 103-114. http://dx.doi.org/10.14257/ijgdc.2018.11.9.09

Buabin, E. (2012). Boosted hybrid recurrent neural classifier for text document classification on the Reuters news text corpus. *International Journal of Machine Learning and Computing*, *2*(5), Article 588. https://dx.doi.org/ 10.7763/IJMLC.2012.V2.195

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modelling. *ArXiv,* Article 1412.3555. https://doi.org/10.48550/arXiv.1412.3555

Dahou, A., Xiong, S., Zhou, J., Haddoud, M. H., & Duan, P. (2016). Word embeddings and convolutional neural network for Arabic sentiment classification. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 2418-2427). The COLING 2016 Organizing Committee.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *Arxiv,* Article 1810.04805. https://doi.org/10.48550/arXiv.1810.04805.

El Mahdaouy, A., Gaussier, E., & El Alaoui, S. O. (2017). Arabic text classification based on word and document embeddings. In A. Hassanien, K. Shaalan, T. Gaber, A. Azar, M. Tolba (Eds.), *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016. Advances in Intelligent Systems and Computing* (Vol. 533). Springer. https://doi.org/10.1007/978-3-319-48308-5_4

El-Alami, F. Z., & El Alaoui, S. O. (2016). An efficient method based on a deep learning approach for Arabic text categorization. *International Arab Conference on Information Technology*, 1-7.

El-Alami, F. Z., El Alaoui, S. O., & En-Nahnahi, N. (2020). Deep neural models and retrofitting for Arabic text categorization. *International Journal of Intelligent Information Technologies (IJIIT)*, *16*(2), 74-86. https://dx.doi.org/10.4018/IJIIT.2020040104

Elfaik, H., & Nfaoui, E. H. (2021). Combining context-aware embeddings and an attentional deep learning model for Arabic affect analysis on Twitter. *IEEE Access*, *9*, 111214-111230. https://doi.org/10.1109/ACCESS.2021.3102087

Elhassan, R., & Ahmed, M. (2015). Arabic text classification review. *International Journal of Computer Science and Software Engineering (IJCSSE), 4*(1), 1-5.

Elnagar, A., Al-Debsi, R., & Einea, O. (2020). Arabic text classification using deep learning models. *Information Processing & Management*, *57*(1), Article 102121. https://doi.org/10.1016/j.ipm.2019.102121

Feng, S., & Kirkley, A. (2021). Integrating online and offline data for crisis management: Online geolocalized emotion, policy response, and local mobility during the COVID crisis. *Scientific Reports*, *11*, Article 8574. https://doi.org/10.1038/s41598-021-88010-3

Galal, M., Madbouly, M. M., & El-Zoghby, A. D. E. L. (2019). Classifying Arabic text using deep learning. *Journal of Theoretical and Applied Information Technology*, *97*(23), 3412-3422.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks,* 2005 (Vol. 4, pp. 2047-2052). IEEE Publishing. https://doi.org/10.1109/IJCNN.2005.1556215

Guzman, E., Alkadhi, R., & Seyff, N. (2017). An exploratory study of Twitter messages about software applications. *Requirements Engineering*, *22*, 387-412. https://doi.org/10.1007/s00766-017-0274-x

Hmeidi, I., Al-Ayyoub, M., Abdulla, N. A., Almodawar, A. A., Abooraig, R., & Mahyoub, N. A. (2015). Automatic Arabic text categorization: A comprehensive comparative study. *Journal of Information Science*, *41*(1), 114-124. https://doi.org/10.1177/0165551514558172

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *ArXiv,* Article 1801.06146. https://doi.org/10.48550/arXiv.1801.06146

Hunt, E. (2016). *What is fake news? How to spot it and what you can do to stop it.* The Guardian. https://www.theguardian.com/media/2016/dec/18/what-is-fake-news-pizzagate

Ibrahim, M. F., Alhakeem, M. A., & Fadhil, N. A. (2021). Evaluation of Naïve Bayes classification in Arabic short text classification. *Al-Mustansiriyah Journal of Science*, *32*(4), 42-50.

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS Transactions On Computers*, *4*(8), 966-974.

IMF. (2016). *Economic diversification in oil-exporting Arab countries*. International Monetary Fund. https://www.imf.org/en/Publications/Policy-Papers/Issues/2016/12/31/Economic-Diversification-in-Oil-Exporting-Arab-Countries-PP5038

Jefferson, H. (2018). *Get old tweets programmatically*. Github. https://github.com/Jefferson-Henrique/GetOldTweets-python

Jordan, S. E., Hovet, S. E., Fung, I. C. H., Liang, H., Fu, K. W., & Tse, Z. T. H. (2018). Using Twitter for public health surveillance from monitoring and prediction to public response. *Data*, *4*(1), Article 6. https://doi.org/10.3390/data4010006

Karali, S. M. Thanoon, C. T. Lin. (2021). Arabic news tweets. *Mendeley Data, V3*. http://dx.doi.org/10.17632/9dxgbgx86k.3

Khoja, Y., Alhadlaq, O., & Alsaif, S. (2017). *Auto Generation of Arabic News Headlines*. Stanford University.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information, 10*(4), Article 150. https://doi.org/10.3390/info10040150

Lehman-Wilzig, S. N., & Seletzky, M. (2010). Hard news, soft news, 'general'news: The necessity and utility of an intermediate classification. *Journalism*, *11*(1), 37-56. https://doi.org/10.1177/1464884909350642

Matrane, Y., Benabbou, F., & Sael, N. (2021). Sentiment analysis through word embedding using AraBERT: Moroccan dialect use case. In *2021 International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA)* (pp. 80-87). IEEE.https://doi.org/10.1109/ICDATA52997.2021.00024

Moh'd Mesleh, A. (2011). Feature subset selection metrics for Arabic text classification. *Pattern Recognition Letters*, *32*(14), 1922-1929. https://doi.org/10.1016/j.patrec.2011.07.010

Mohammed, P., Eid, Y., Badawy, M., & Hassan, A. (2020). Evaluation of different sarcasm detection models for Arabic news headlines. In A. Hassanien, K. Shaalan, & M. Tolba (Eds.), *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2019. Advances in Intelligent Systems and Computing,* (Vol 1058). Springer. https://doi.org/10.1007/978-3-030-31129-2_38

Panagiotou, N., Katakis, I., & Gunopulos, D. (2016). Detecting events in online social networks: Definitions, trends and challenges. In S. Michaelis, N. Piatkowski, & M. Stolpe, M. (Eds.), *Solving Large Scale Learning Tasks. Challenges and Algorithms. Lecture Notes in Computer Science*, (Vol 9580). Springer. https://doi.org/10.1007/978-3-319-41706-6_2

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *ArXiv,* Article 1802.05365. https://doi.org/10.48550/arXiv.1802.05365

Raftery, T. (2017). *Twitter Arab Word - Statistics Feb 2017*. https://weedoo.tech/ twitter-arab-world-statistics-feb-2017

Rey, M. V. (2019). *What are the eleven parts and their meaning*. Philippine News. https://philnews.ph/2019/07/16/parts-of-newspaper/#google_vignette

Saeed, M. (2021). Farasapy: *A Python wrapper for the well Farasa toolkit.* Github. https://github.com/MagedSaeed/farasapy

Salloum, S. A., Al-Emran, M., & Shaalan, K. (2017a). Mining text in news channels: A case study from Facebook. *International Journal of Information Technology and Language Studies*, *1*(1), 1-9.

Salloum, S. A., Al-Emran, M., & Shaalan, K. (2017b). Mining social media text: Extracting knowledge from Facebook. *International Journal of Computing and Digital Systems*, *6*(02), 73-81. http://dx.doi.org/10.12785/IJCDS/060203

Salloum, S. A., Al-Emran, M., Monem, A. A., & Shaalan, K. (2018). Using text mining techniques for extracting information from research articles. In K. Shaalan, A. Hassanien, A., & F. Tolba (Eds.), *Intelligent Natural Language Processing: Trends and Applications. Studies in Computational Intelligence*, (Vol 740). Springer. https://doi.org/10.1007/978-3-319-67056-0_18

Sayed, M., Salem, R., & Khedr, A. E. (2017). Accuracy evaluation of Arabic text classification. In *2017 12th International Conference on Computer Engineering and Systems (ICCES)* (pp. 365-370). IEEE Publishing. https://doi.org/10.1109/ICCES.2017.8275333

Schmidhuber, J., & Hochreiter, S. (1997). Long short-term memory. *Neural Comput*, *9*(8), 1735-1780.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (pp. 1-11). NeurIPS Proceedings.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. ICML '97: Proceedings of the European Fourteenth International Conference on Machine Learning (pp. 412 - 420), Morgan Kaufmann Publishers Inc. https://surdeanu.cs.arizona.edu/mihai/teaching/ista555-fall13/readings/yang97comparative.pdf.