National College of Ireland

Online News Popularity Prediction using LSTM and Bi-LSTM

M.Sc. Research Project Report
M.Sc. in Data Analytics

Prasad Rudrappa Shivu
Student ID: X19213077

School of Computing
National College of Ireland

Supervisor: Prof. Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Prasad Rudrappa Shivu
Student ID:	X19213077
Programme:	M.Sc. in Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Prof. Jorge Basilio
Submission Due Date:	16/08/2021
Project Title:	Online News Popularity Prediction using LSTM and Bi-LSTM
Word Count:	7014
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard speci ed in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Prasad Rudrappa Shivu
Date:	16/08/2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator o ce must be placed into the assignment box located outside the o ce.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Online News Popularity Prediction using LSTM and Bi-LSTM

Prasad Rudrappa Shivu
X19213077

Abstract

Online news has become an apparently essential source of information, capturing the attention of a considerable audience. People are reading and posting news more often online, on websites, blogs, and social media sites such as Facebook and Twitter. The rapid advancement of smart phones, as well as the introduction of other electronic devices, has made access to news far simpler than ever before. Every day, millions of news stories are published online. The writings are short-lived; some fade away quickly, some acquire traction as their reach grows, and others reclaim lost attention thanks to follow-up news or related articles. Predicting their popularity would benefit not just news stakeholders, but also the general public, who would be kept up to speed with the newest information. Several studies have been conducted over the last decade to forecast the popularity of news using various machine learning algorithms. While some researchers utilized factors such as the number of times the news was shared on social media, the amount of visitor comments, and the overall number of likes the article received to forecast popularity, others relied on word count, title, and abstract length. To forecast popularity, a variety of regression and classification techniques were used. The models' accuracy, on the other hand, appeared to be approximately 70%. This research uses LSTM and Bi-LSTM, two powerful unsupervised machine learning models, to forecast the popularity of news items before they are published, utilizing Feature Selection to increase the model's accuracy. The accuracy of the model improved to 79% for Bi-LSTM with normally distributed features with high Fisher Score. Aforementioned model, showing improved accuracy, can be deployed to predict the popularity of news published across online news media and social media.

1 Introduction

Several businesses are determined to gure out how much demand for the news pieces they produce will grow in the future. As a consequence, they're ready to make quick judgments to put their ideas into action on whichever platform is accessible (Namous et al.; 2018). It's critical to have enough resources (recent news) to support the market's disruptive needs. Subject, substance, timing, article placement on the web page, language, resemblance to current events, past success of comparable subjects, time since news release, popularity season, and connection to ongoing popular events are all elements that impact popularity prediction. Furthermore, internet news has a number of qualities that make it more appealing to news organizations, such as small content sizes, inexpensive publication costs, and simple wording (Tatar et al.; 2012). If a piece of news

is expected to be popular, the company might plan to increase its appeal and timeliness in order to maximize income (Xiong et al.; 2021). Furthermore, by streaming advertising or showing banners with news stories, they may provide maximum visibility to ad clients.

A steady stream of news items makes it simple to integrate with social networking sites. According to the Pew Research Center, nearly 69 percent of adults in the United States use Facebook (FB) (Auxier and Anderson; 2021), and more than "eight-in-ten" of them (86%) bank on FB for news and updates. Television advertisements, movie sales calculations, automobile tra c management, and economic trend forecasts are all examples of applications where popularity predicting is bene cial. Companies spend well over 30% of their money on internet marketing (Tatar et al.; 2014). This has intensi ed rivalry among online websites and social media pages to not only captivate the audience through the quality of news they give, but also to bring in well-run businesses as advertising clients. Furthermore, news sites might utilize forecasts to choose their most important items and organize their homepage accordingly. Liu et al. (2017), to entice readers by appealing to their topics of interest Furthermore, those who read on the internet may lter information as needed and concentrate on the essentials (Tatar et al.; 2012). On the other side, it may assist authorities in preventing the dissemination of false information. Given the numerous advantages of forecasting the popularity of internet news, it is critical to be aware of the research that has been done thus far and to add more to the domain. Section 2 discusses previous contributions to the issue under investigation.

Predicting news popularity will allow sites to optimize their news snippets based on the most current trend by determining whether or not they will engage readers. As a result, utilizing business intelligence and machine learning models, an optimal automated prediction model must be built. Because predicting popularity after publishing appears to be very simple when evaluating the actual information provided, the number of likes and comments received after it is made available online, the prediction accuracy is high. Furthermore, post-publication prediction follows the zeitgeist of utilizing news-reach as a predictor of popularity. Anticipating how popular it could be is still di cult since metadata features are taken into account rather than real content (Uddin et al.; 2016). Furthermore, popularity is totally dependent on user behavior at any one time, as well as interest, sentiments, and sentiment toward the issue, all of which are di cult to anticipate. Many impartial aspects are lacking in prediction. Given that reach is in uenced by a variety of factors, these unrelated variables may be useful in forecasting popularity. Furthermore, predicting at the microscopic level is di cult due to the changing nexus of social interactions and information cascades.

Over the last decade tremendous attempts has been made to predict the popularity of online news. While most proposed models were for predicting popularity post publication, a few tried predicting the reach of articles prior publication by making convenient assumptions. The research work can be seen in Section 2.4. However, the accuracy of the models still remain low. This encouraged further research in the eld of online news prediction. This paper aims at predicting the popularity of news published online using unsupervised learning algorithms - LSTM and Bi-LSTM, with a motive to provide improved answers through deep learning models.

Section 2 debriefs on the related work done over time. Section 2.2 briefs on the work done on predicting popularity of news considering Twitter as the source. Section 2.3 discusses work done on YouTube videos popularity. Section 2.4 discusses models employed by researchers in the eld of study, the results of which are described in 2.5 under section 2.5. Section 3 gives the overview of the methodology followed in the research

. Data acquisition, data description, exploratory data analysis, data transformation and models implemented to investigate the dynamics of articles popularity are detailed in section 4. Subsequently, the results and evaluation methods used to compare the results of the proposed models are debriefed and tabulated in section 5. Finally, the work is concluded and future work is suggested in Section 6.

2 Related Work

2.1 Introduction

The ubiquity of the Internet has resulted in a plethora of computer-mediated communication channels. Because internet chat has become such an essential component of modern society's communication, interest in it is continuing to increase at a rapid pace. Sequential behavior, which is observed as a pattern of individuals responding to articles or comments, has established the discussion threads and resulted in the tree network structure. The gathering and analysis of data on the internet provides intriguing insights into human behavior. Statistically and theoretically justi ed models are required to establish which social variables are accountable for network topologies, despite assertions that data-driven approaches would render the scienti c method obsolete.

In recent years, research has o ered many modelling approaches for determining structures by extracting various characteristics linked to human behavior. As a result, they're also known as generative models, because they not only compute the statistical signi cance of functions, but also reconstruct the next forthcoming communications patterns of any comment chain. The literature review provided here brie y describes various studies conducted on online news, covering various topics such as active learning, implementing machine learning, and deep learning models to predict and evaluate research results, achievements, and suggested future work. Further study of the topic provides a gateway to answer questions of interest on the topic.

2.2 Twitter as a source of online news articles

New digital social networks such as Twitter have developed as a dominating platform for individuals to connect and exchange knowledge, news, and viewpoints as social media sites have grown in popularity. These technologies seem to have the ability to change existing social structures and connections.

Housley et al. (2018) showed interest in learning more about user behavior and interaction with social media material. To emphasize the signi cance of an interactional approach to social media research, a comprehensive study of Twitter-based online campaigns was undertaken. After categorizing social media posts based on activities, a typology of user interactions was established. The necessary tweets were collected using the COSMOS program, which enables for automated data processing.

Aragon, Gomez and Kaltenbrunner (2017) addressed two key issues: a lack of data or a model that is insu cient, and the uncertainty of the social structure. To inspire and explain issues, a simple stylized model of success was created, which interprets the success of any knowledge based on its competence. The model was used to investigate twitter cascades by de ning the tweet dataset, extracting features, and comparing the performance of each model using Linear Regression, Random Forest, and Regression Trees. The tweets in the dataset are from February 2015 to April 2017. To evaluate the

genuine tweets pertaining to the photos and articles, a social media classi er was used to lter out tweets with a high spam count, and then tweets from popular domains were considered over time. The features were retrieved from the content, which included the domain, tweet time, spam score, and category, as well as from the people, who included followers, friends, and statuses, as well as subject features. They are coarse-grained, missing information that distinguishes users from material, as well as their interaction. As a result, for more advanced users, a standard topic model was created. The dataset was divided into training and test groups and then processed using the methods described above. Random forest regularly delivered correct results and deduced the ex-ante prediction's limits. When it comes to social media, news stories have had a signi cant in uence due to the variety of information tactics used.

The news items, according to Bandari et al. (2012) are extremely time sensitive. Furthermore, news pieces are in a strong battle to disseminate as much as possible. As a result, predicting the popularity of news items on social media is both intriguing and di cult. The classi cation and regression algorithms were used to determine the internet popularity. The dataset is made up of two pieces of information: rst, published news items from the web for a speci ed time period, and second, articles posted on Twitter by the user. The data was gathered using Feedzilla, a news feed aggregator API, and Topsy, a twitter search engine, and the number of each item linked on twitter was calculated. After that, the data is pre-processed to eliminate spam tweets, duplicates, and name variants. The characteristics were based on the type of news on social media, well-known names in news stories, language attachment, and the person who published the news piece that mostly covered the material. The results indicated that the leading source of information on Twitter was not from the top news organizations, but rather from technology blogs such as Google blogs and many others, with an accuracy of 84 percent. The stories become extremely popular on the internet as a result of their widespread distribution, in uencing middle-class individuals to raise awareness and guarantee that the news reaches the intended audience.

According to research work by Rathord et al. (2019), interactions on social media could reveal remarkably accurate forecasts about the future. It is shown that utilizing online credibility as an indicator of social acceptance based on Twitter content, it is possible to forecast new venture survival with a high degree of accuracy. Over 187,000 tweets from 253 new enterprises' Twitter accounts were analyzed using context-speci c machine learning. Because potential customers sometimes lack a frame of reference for comprehending the bene ts of a new venture's product, a company's level of expertise about its goods, management, and organization is seen as a key source of validity. The quality of such posts has been found to be a good determinant of a company's online trustworthiness in studies. Because credibility, an organization's intangible but most valuable commodity, is linked to legitimacy, which is de ned as an individual's view of an institution. The information was received from a large nancial rm located in Switzerland. The 253-entry dataset comprises primarily of names, social media accounts and web pages for companies, and the number of enterprises. The data was extracted using the Twitter REST API and included 187,323 tweets, 441,583 likes, and 102,501 retweets from June 2018. Three separate aspects of online legitimacy were examined: information quantity, information substance, and interaction and con rmation content. On the dataset, the methods Random Forest and Gradient Boost were used. The ndings revealed a 76 percent accuracy, indicating that a ve-year forecast could be made with high precision and recall values.

For obtaining hyper parameter, partitioning the data set, training the classi cation model, applying the model, and assessing the outcomes, Wicaksono and Supianto (2018) presented a genetic method. The ndings from the Genetic Algorithm and Grid Search were virtually identical. However, in terms of processing time, the Genetic Algorithm performed well.

Sentiment analysis has exploded in popularity, with the objective of utilizing machine-learning techniques like sentiment, subjectivity analysis, and polarity calculations to investigate the opinions or content on various social networking platforms. Despite the adoption of a variety of machine-learning algorithms and methodologies for sentiment analysis during elections, a statistic was urgently required. Nave Bayes and SVM were used in the analysis. The information acquired through the Tweepy API was connected to accounts. The hashtags chosen were connected to prominent political viewpoints on Twitter. Only tweets in English and Urdu languages were included in the data, which was pre-processed to eliminate undesirable words, symbols, and vacant spaces. The remaining Urdu tweets were translated to English and discarded. SentiWordNet, W-WSD, and TextBlob sentiment lexicons were used to determine the sentiments.

Similarly, Glowacki et al. (2018) employed Twitter as a medium to collect political data for sentiment analysis during the presidential elections in Mexico and Brazil. Twitter was also used to make connections with people by exchanging academic knowledge.

Machado et al. (2015) performed a survey to determine the importance of academics in the discovery of information and dissemination of knowledge.

2.3 Popularity of YouTube as a medium in sharing data

YouTube is one of the most widely used platforms for sharing information. Vlog has gained a big following and has turned out to be the most popular form of digital entertainment Carral et al. (2019). It delivers profound contextual knowledge as well as enjoyment, and as a result, millions of people watch it.

Ananda and Sandi (2019) used Hermeneutic Phenomenon as the study design to look into the usage of technology in learning English. It is not just a platform for sharing videos, but also a strong medium for online news discussions and political campaigns, since it is the world's second most popular website. It also intends to investigate the political communication of video sharing by performing qualitative research on the Russian presidential election Litvinenko (2021). This was done on 169 videos gathered over the course of two months. The method was taken a step further by performing computational analysis in addition to the quantitative analysis conducted on the YouTube videos.

A lifetime aware regression model was proposed by Ma et al. (2017) who worked on two datasets, one of which was structured around hourly data collection. There were 172,602 videos in all. The other was on a daily basis, and it has 631,459 videos in it. For 100 days, the popularity of YouTube videos was recorded and analyzed to determine what patterns and trends emerged. LARM's unique feature is that it divides the dataset into numerous subgroups and trains each subset independently, demonstrating linearity for both currently observed and future popular videos. The model was matched with other e cient models such as the MRBF (an extension of the MR model), the View Count Dynamic Model (VCDM), and the Regression Tree, and a lifespan metric called αLifespan metric was devised. Following a comprehensive comparison with other models, LARM was projected to lower the prediction error for hourly and daily data by 18%.

2.4 Implementation of machine learning and deep learning models on news articles

The internet allows digital news to be instantly spread throughout the world. The majority of people today consume and share news on the internet, for example, through social media platforms like Twitter and Facebook. The amount of readers, likes, and shares on a piece of news is typically a reliable measure of its popularity. It is extremely bene cial to anyone involved in internet news, such as content producers and advertisers. As a result, applying machine learning to predict the performance of online news stories is both intriguing and practical. Several studies have been carried out in order to predict the success of internet news. The popularity of news is determined by a number of factors, including the posting of online news on social media, visitor feedback on news, likes for new posts, and so on. It is therefore critical to understand, analyze, and predict the pattern in order to provide successful news articles that reach the greatest number of people.

Growing interest towards social media has contributed to the enhancement of statistical models. Social media has provided the major platform for publishing comments, which provides a channel for online debate. Aragon, Gomez and Kaltenbrunner (2017) studied statistical modeling of online conversations in depth, with an emphasis on existing generative models of discussion thread construction and evolution. The goal of building generative models and statistically assessing them using empirical evidence is to characterize the processes that govern online debate dynamics. These are parameterized network construction models that might be used to generate synthetic discussion threads that resemble some of the features of genuine conversations.

Optimizing a likelihood function, which assesses how well the model explains the data as a function of its parameters, is the most common approach for estimating model parameters. Harrison et al. (2018). Although for basic models this optimization may be done analytically, it can be computationally expensive in general. The complexity of the model determines the challenge of such an optimization. When compared to the amount of the data, a model with numerous parameters cannot be generalized; rather, the data can be split and regularized into distinct subsets for training, testing, and validation.

By analyzing features accessible on platforms, predicting user behavior, and evaluating platform design, researchers Aragon, Gomez, Garc a and Kaltenbrunner (2017) were able to analyze online discussion threads across di erent platforms and user behavior, as well as assess the in uence of design on online platforms. Nashaat and Miller (2021) ArtAI, an enhanced prediction model that feeds news articles as input, was proposed. The size of the data collection, characteristics, popularity metrics, and popular articles are all included in each dataset. Gradient Boost, Vector Space Model, and Ensemble Models are used as foundation models to compare the models. Di erent performance metrics, such as Accuracy, Precision, Recall, and F1 Score, were used to assess the outcome. The experimental ndings indicated that by using a meta-active learning method, ArtAI may achieve greater accuracy and e ciency than other base models. It also resulted in high MCC values for all tasks, leading to the conclusion that the suggested model encourages users to correct errors in order to deliver relevant news items to consumers, and that the proposed strategy might improve state-of-the-art performance by 19.72 percent. In comparison to traditional active learning, it also had a 37.09 percent better outcome.

Di erent techniques were used by Rathord et al. (2019) to forecast internet news stories. Each article was distinguished and described by 59 characteristics in the dataset. The

features are chosen with the least amount of repetition and the greatest amount of relevance in mind. Several machine learning models were used to process the data, including Random Forest, Adaptive Boosting, Support Vector Machine, K-Nearest Neighbors, and Nave Bayes, Linear Regression, Logistic Regression. Di erent performance matrices, including as accuracy, precision, recall, F-score, ROC Curve, and AUC, were used to assess the results. The results revealed that Random Forest performed the best and that there was room for development in terms of accuracy and optimization for increased popularity. People utilize internet news as a source of fresh information; there are several online news sites, and many people only read news that interests them. This sort of news is common and bene cial for media owners. As a consequence, it is feasible to forecast whether a news item will become popular or not using prediction algorithms. Machine learning is one of the most commonly utilized approaches. To enhance prediction accuracy, the optimum hyper parameter of machine learning algorithms must be computed. Because grid search attempts every conceivable combination of hyper parameter, determining the hyper parameter with this approach might take a long time. This is an issue since we need more time to forecast the popularity of internet news.

A Comprehensive research was carried out in this regard by Namous et al. (2018). A neutral network, as well as other machine learning models, were used to identify and analyze online popularity. The use of several models revealed that SVM and neural networks provided an overall accuracy of 65 percent. The widespread usage of smart phones has also enhanced the appeal of internet news by allowing people to rely on it.

(Obiedat; 2020) Random Forest, Logistic Function, Bayes Net, Simple Cart, and C4.5 are some of the machine learning models that have been introduced for prediction. Kappa statistics, TP-Rate, Root Mean Square Error, Accuracy, Precision, F-Measure, and FP-Rate are used to assess the measures' performance. To increase performance, data is processed using feature ltering techniques, which follows a straightforward approach that begins with data collection, then moves on to model construction and evaluation. The model was built by Fernandes et al. (2018). For evaluating the aforementioned and other models, Kiranmai and Ahuja (2018) installed Weka 3.9.3 was on windows 10 OS which includes 8 GB RAM. It's a popular open-source machine learning platform created by a team in New Zealand using JAVA. Classi cation, Clustering, Regression, Feature Selection, and Visualization are examples of data pre-processing and data mining operations that might bene t from the aforementioned paradigm. The data is split into ten subgroups, with one set used for testing and the others for training. The confusion matrix was used to assess the results.

Study on the same dataset by Haghighi et al. (2018) had four binary classi ers like { True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The feature selection technique used the Weka tool, which assists in the removal of super uous features in order to improve accuracy. The dataset was being applied with ve di erent selection methods like:

1. **InfoGainAttributeEval:** Determines the attribute's relation to the designated class using the attribute's knowledge gain as the key metric.

2. **ChiSquaredAttributeEval:** Calculates the Chi-Squared statistic for the attribute as a key predictor of its relevance to the designated class.

3. **CorrelationAttributeEval:** Employs the Pearson's Correlation with class label as the key indicator of an attribute's relevance in relation to the labelled class.

4. **GainRatio:** Calculates the bene t ratio with respect to the class to nd the value of an variable.

5. **OneRAttributeEval:** One-R is a basic algorithm (used by Riganello et al. (2018) to nd out disorders using Heart Rate Variability Entropy) that determines one rule in the training data and then chooses the rule with the least error for all attribute. It can deal with missing values by considering "missing" to be a valid value.

Despite the fact that it is a reduced form of the classi er, it might be useful for assessing and establishing a benchmark for other learning methods. Random Forest with the parameter InfoGainAttributeEval had the highest accuracy of 66.9% and a Kappa Statistic value of 0.372 of all the models implemented. Other performance matrices, such as Precision, Recall, and F-Measure, all indicated a value of about 67 percent, with no room for improvement. One of the most important aspects of this project was the use of feature selection procedures, which allowed the authors to concentrate on the key themes and articles that gained greater popularity and attracted a big audience. The forecast became more speci c as the study in internet news continued, and the analysis was limited to a speci c area of interest. Assume that your eld of interest is cryptocurrency (Beck et al.; 2019), and your study is solely focused on news about it. Real-time bitcoin news was gathered from the web and compared to social media tweets. Di erent algorithms were employed to forecast the articles referenced on Twitter in the next 24 hours based on the day tweet activity. The information was gathered in three steps: gathering data from Twitter, gathering articles, and matching tweets to articles. The real-time data was gathered using the Twitter API and ltered using a list of cryptocurrency related phrases. The articles were gathered from Gazetter source URLs, with a total of 2353 articles, and articles were matched with tweets using a document-oriented database that explicitly connects the tweet with the article's URL. By selecting three below listed requirements, the various entries of the article's URL are combined into a single URL.

1. If the URLs have the same host and path

2. If the heading of the URLs match

3. If the articles are published at the same time

With a training set of 125248 articles and Auto-regression, Random Forest, and sequence to sequence models as the baseline for the study, the researchers utilized three categories of characteristics: time series features, content features, and context features. The outcome was assessed using two di erent sorts of characteristics.

1. **Mean Absolute Percentage Error:** This estimates the average percent di erence between the predicted and actual values. The reason for utilizing percentage mistakes rather than absolute errors is the huge disparity in the number of articles listed. An absolute error metric would very de nitely be dominated by a few publications with a lot more mentions.

2. **Normalized Discounted Cumulative Gain:** The second fascinating aspect of the predictions is the sequencing of the articles as a result of the projected values. We could obtain the most intriguing articles from a model that has a low MAPE but generates a good approximation of the news ordering. The Discounted Cumulative Bene t (DCG) is high if the top projected articles receive a lot of attention.

The results indicated that all of the models performed well; however, articles published after 15 to 20 hours performed better, with a 20% increase or decrease in MAPE value and an NDCG value of 0.9 reached after 5 hours of publication. S2S performed the worst out of all the models. Even though the system provided accurate results, it might be improved by training the model to analyze popularity after an hour of publishing and using sophisticated NLP models.

Similarly, as the study continues, journalists and editors are keen to learn about the pieces that have been published as quickly as possible, as well as the ones that bring the greatest tra c to their websites. The linked study includes a range of approaches and algorithms for conducting this forecasting. Before making any longer-term predictions, most of the proposed approaches necessitate tracking the success of material for a period after it is released. A new method for forecasting the success of news stories published on the internet; the method is based on a variety of assumptions about article similarities and topicality, and it complements current content-based approaches. To begin with, the popularity of a new article was correlated with the popularity of comparable pieces that had previously been published. Second, the success of the new piece was attributed to the primary topic's recent historical popularity. Based on these data, we utilize time series forecasting to predict the amount of visitors an item will receive. The information was gathered from a worldwide news network with many channels and websites. The data was collected between 2012 and 2014 and consisted of two types of articles: breaking news and issues that occurred in various regions or throughout the world, as well as reporting on the activities. Second, the remarks or thoughts expressed by well-known authors and editors on any public issue or item. Latent Dirichlet Allocation (LDA) models, as well as machine learning methods such as linear regression and Support Vector Regression models, were used. Pearson's Correlation was employed to assess the auto correlation and cross correlation in the themes in order to forecast the outcomes. To begin, it was determined that the themes selected were content-coherent. Second, there were auto-correlations in the topic volume time series. Predicting the success of an article requires a combination of content-based approaches that evaluate the communicative frame of the articles and time series methods that capture the growth of people's attention around speci c subjects.

The technique e ectively incorporated two factors in anticipating article visits: the popularity of similar articles in the most recent issue, as well as the popularity of the themes addressed in the article. More precisely, it was demonstrated that when those two variables were combined, the model outperformed each component alone. Furthermore, when subject predictions are merged, which can be done with as low as 2.5 percent inaccuracy, the total mean average error rate is around 11%.

Almost all latest research work, during past 3-4 years, emphesises on quantitative analysis. The research by Nashaat and Miller (2021) coined a model called ArtAI and matched the result with other ML models such as Gradient Boost, Vector Space models.

Ma et al. (2017), Rathord et al. (2019) and Wicaksono and Supianto (2018) worked on the data using ML models such as Support Vector Machine, Linear Regression, Gradient Boost, Random Forest and other models. The work revolved around extracting the features those unmatched the features in other research, as it all depends on the features that are primarily focused in the above cited mentioned research works.

2.5 Results

Table 1: Summary details of the related work on Online News Popularity Prediction

Author(s)	Objectives	Research Design	Algorithms and Accuracy	Findings
Namous et al. (2018)	Employed multiple ML models to nd the popularity of online news	Quantitative evaluation of di erent algorithms	MultiLayer Perceptorn Bagging AdaBoost Random Forest Naive Bayes K Nearest Neighbour Logistic Regression Linear SVM Polynomial SVM RPF SVM Sigmoid SVM	65% 64% 64.8% **65.8% - Best Result** 64.7% 57.5% 55.9% 48.3% 46.8% 53.5% 51%
Ren and Yang (2015)	Reviewed 10 ML algorithms to nd the best model and set of features to predict the popularity of online news	The Performance of the models are recorded and compared. Feature Selection is used for better performance	Linear Regression Logistic Regression SVM (d = 9 Poly Kernel) Random Forest (500 Trees) KNN(k=5) Linear SVR REPTree Kernel Partial Least Square Kernel Perceptron C4.5 Algorithm	66% 66% 55% **69% - Best Result** 56% 52% 67% 58% 45% 58%
Guan et al. (2017)	Implemented Intelligent Decision Support System (IDSS) using 5 Models that analyzes articles prior to their publication	Compare the accuracy of the models to nd the best t	Random Forest AdaBoost SVM KNN Navie Bayes	**73% - Best Result** 72% 71% 67% 65%
Sitapara et al. (2018)	Four ML algorithms were used to predict the online news popularity	Compared the accuracy of the models	Decision Tree KNN Random Forest Naive Bayes	57.9% 57% **66.4 - Best Result%** 61.2%

From the table it is evident that Random Forest is the best proven method so far with an accuracy of 73%.

3 Methodology

This research employs the `Cross Industry Standard Process for Data Mining' (CRISP-DM) methodology. It is a hierarchical process which mainly has six level breakdowns, in the work by Chapman and Wirth (2000). The model however is slightly modi ed to hold good with the work done here. Figure below shows the actual and modi ed versions.

The ow initiates with extracting the data and pre-processing, followed by extracting the suitable feature Extraction. Then the models - LSTM (with and without auto encoding-decoding) and Bi-LSTM (with and without auto encoding-decoding) are Implemented, followed by evaluating the results. Finally, the results are compared with that in the work done earlier to suggest the best t model to predict the online news popularity.

Data for the work is extracted from UCI Machine Learning Repository [1], an open source repository serving machine learning community. The dataset contains the details of the articles published by Mashable [2] in a period of two years. The data is originally

[1] https://archive.ics.uci.edu/ml/datasets/Online+News+Popularity

[2] https://www.mashable.com

acquired and preprocessed by Fernandes et al. (2015), who go on to deep dive on the data collection process.

The data on UCI library is usually clean and require no pre-processing. However, there is a subtle clean up necessary with respect to the column headers. The data is then loaded and fed for exploratory analysis, carried out using Python.

Feature selection is employed to extract features that could better train the model for predicting the popularity of news articles. Here six data subsets are picked based on Fisher Score and Normal Distribution. After splitting, counter() function - a class in collections library, is used to nd the number of data elements in each class. The result showed a distinguished class imbalance.To overcome the class imbalance, a balanced synthetic data was generated using SMOTE technique.

Finally, the new balanced data subsets are trained and tested using four models - LSTM, Auto Encoder-Decoder LSTM, Bi-LSTM, and Auto Encoder-Decoder Bi-LSTM.

The accuracy of the models are compared to results obtained from the earlier research works found in section 2.5 and the best t model is determined.

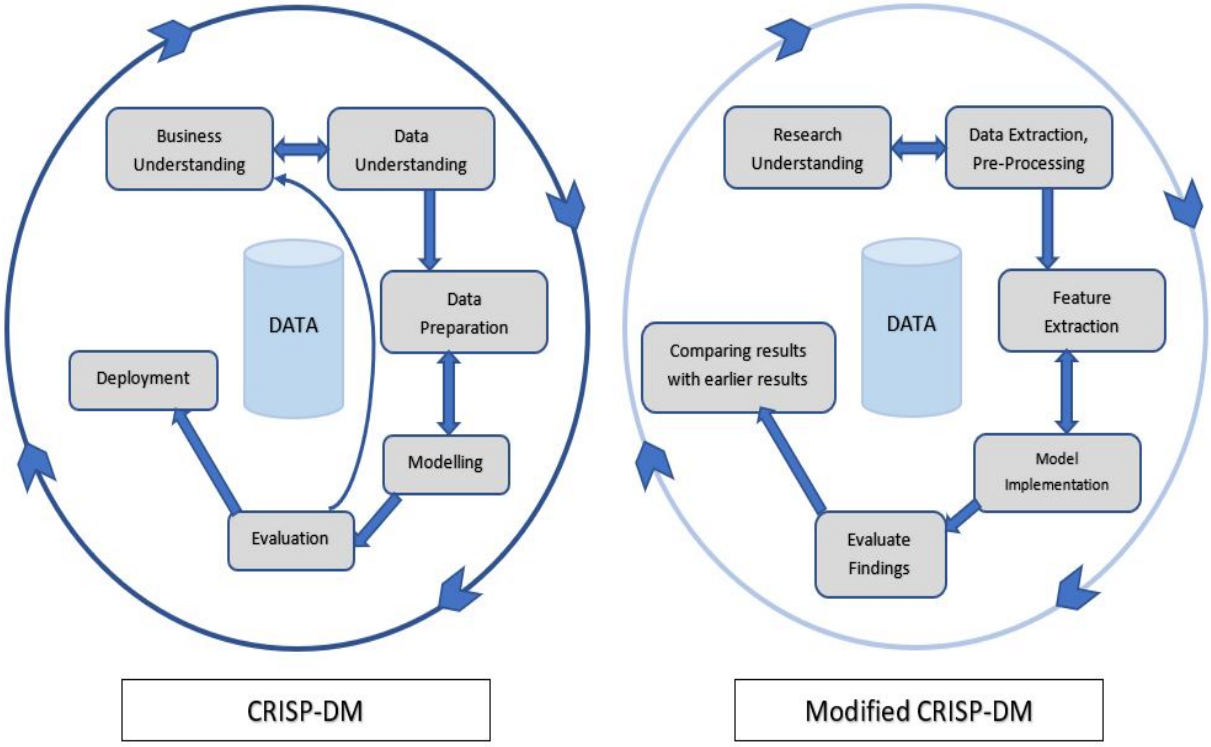

Figure 1: CRISP-DM

4 Implementation

The overview of the implementation can be seen in gure 2. The programming language used is python.

The dataset obtained from UCI Machine Learning Repository has 59 numerical attributes describing various aspects of 39,644 articles. As the data was cleaned extensively (Fernandes et al.; 2015), there is no much pre-processing required. However, the only

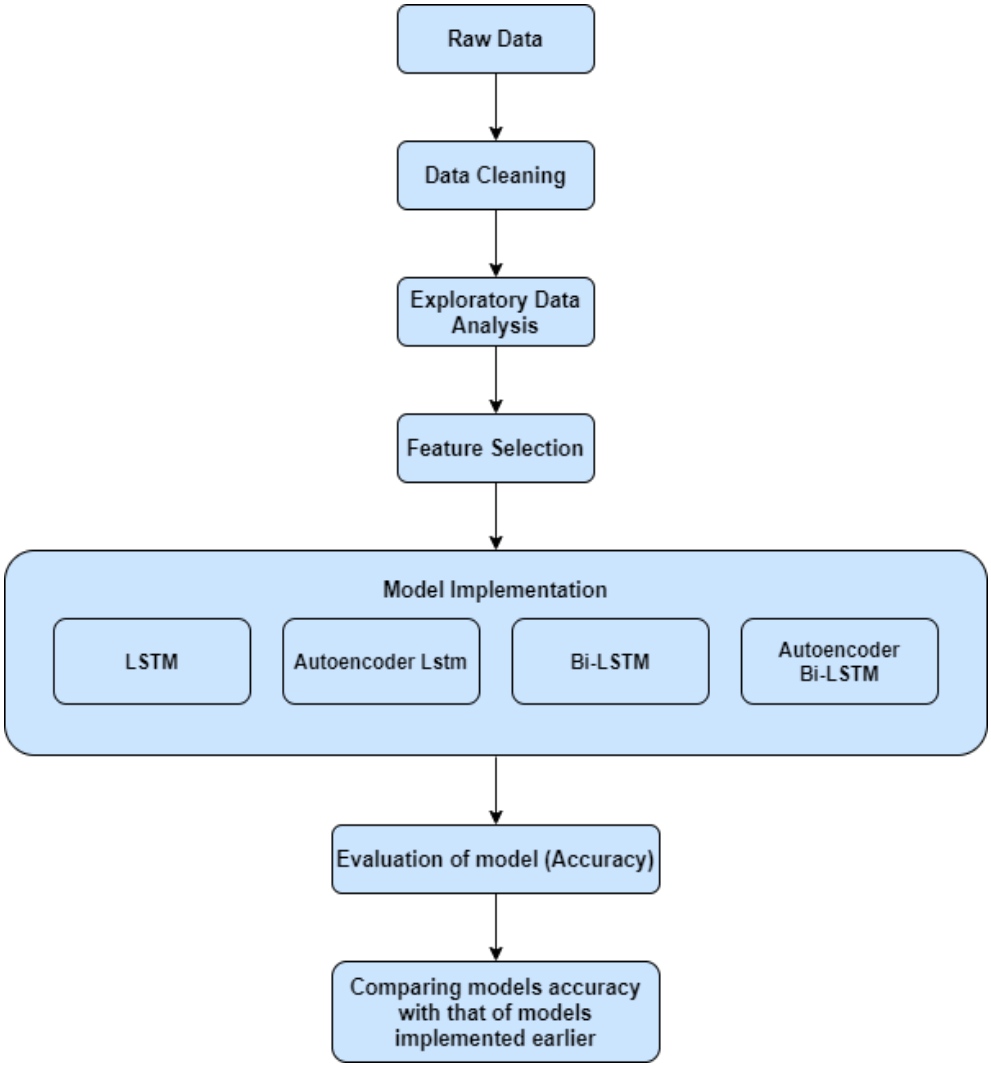

Figure 2: Data Flow of the Research

area that requires cleaning is the headers or column names that have a space appended to the left. This is overcome by improvising while extracting the columns - instead of using 'df['shares']', use 'df[' shares']'. This may appear insigni cant, yet it might lead to some annoying mistakes throughout the execution of the tests. Columns 'url' and 'timedelta' are dropped as they are insigni cant, the number of shares are bucketed into seven categories - Very Poor, Poor, Average, Good, Very Good, Excellent and Exceptional, 'weekdays' columns are merged into one, 'data_channel' columns are merged into a single column, and nally the old data is dropped. The column 'n_non_stop_words' can be classi ed as a noise as most of the records are 0, and can therefore be dropped.

Figure 3 infers, though not always, but frequently popular articles tend of have more reach. Business channels show no in uence with change in num_images. They usually have lesser images irrespective of popularity. Entertainment and tech tends to have more images with increase in popularity.

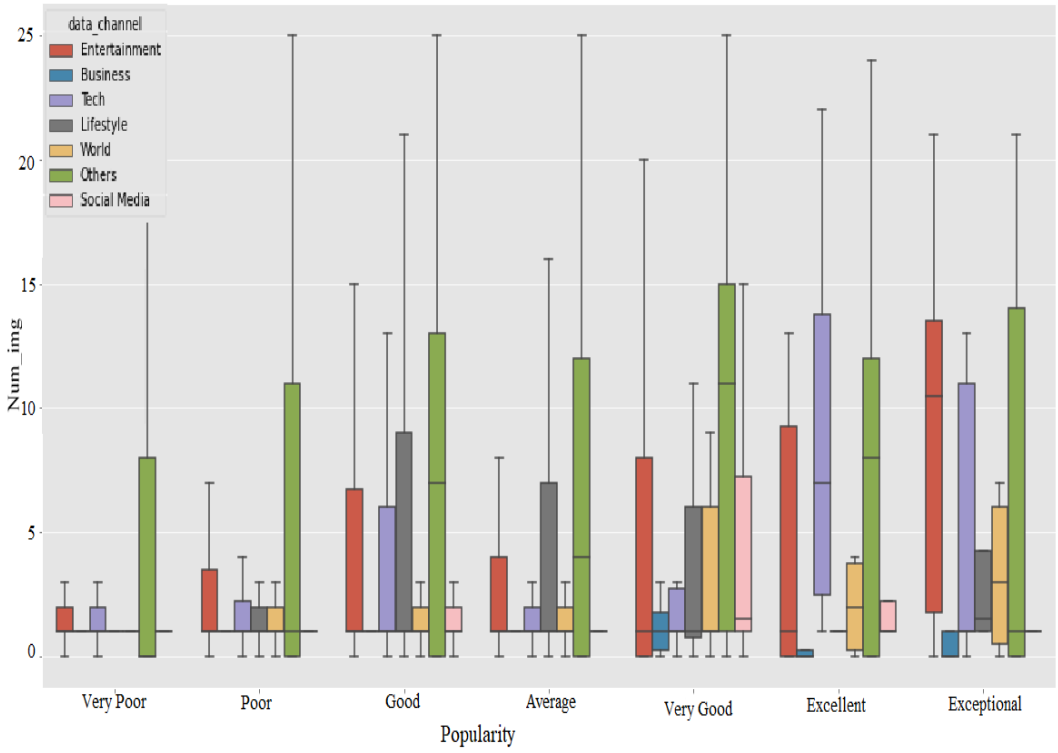

Figure 3: Popularity by Data Channel

Box and Whisker plot in gure 4 shows that most data points lie between 0.6 and 0.8, irrespective of shares.

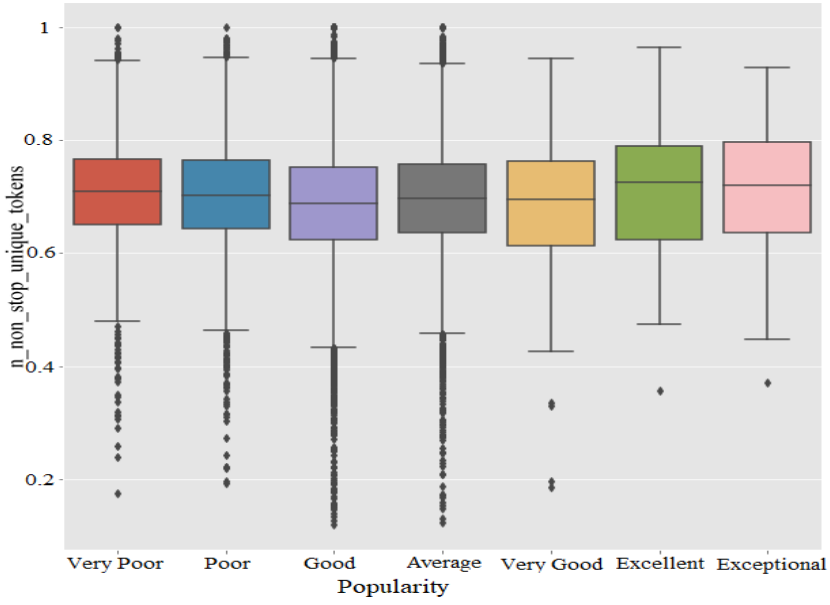

Figure 4: Popularity by Rate of Non-stop words in Content

From gure 5, while both world and tech news are over 7000, social media and lifestyle related news are least with a little over 2000.

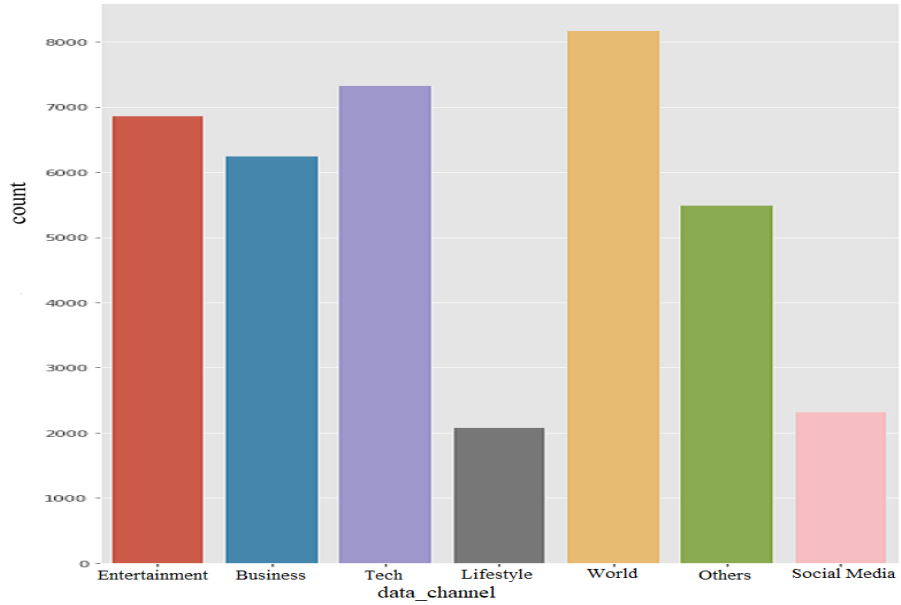

Figure 5: Distribution of Articles across Data Channels

Figure 6 shows that shares doesn't follow normal distribution, instead positively skewed. So, log transformation was applied. The transformed data showed normal distribution and point lie along the diagonal line.

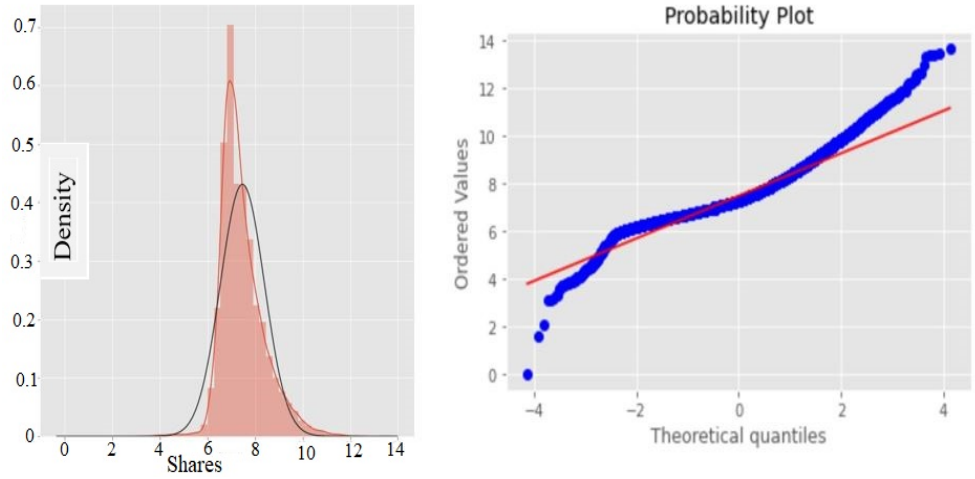

Figure 6: Normality Plot of Shares

As part of feature selection, six set of features are selected and used to train and test the models used in the study.

- First set contains the features on hypothesis - 32 in number

  Hypothesis: Each variable is subjectively analysed to understand how relevant it is to the problem. The features are carefully analysed through univariate and bivariate analysis to accept or reject.

- Second set has all the features in the overall dataset - 58
- Third set has the features with their scores over 3000 - 25 features

- Fourth set contains all the features on hypothesis with normal distribution - 32 features
- Fifth set contains features with fisher score over 100 and show normal distribution - 29 features
- Final set contains all features showing normal distribution - 58 features

The data is split at 70/30. Upon splitting there was class imbalance. Because most machine learning algorithms for classification were built with the assumption of an equal number of instances for each class, imbalanced classifications offer a problem for predictive modelling. As a result, models with poor prediction accuracy, particularly for the minority class, emerge. This is an issue since the minority class is usually the most important. As a result, the minority class is more vulnerable and prone to classification mistakes than the dominant class. The data set here showed a sever imbalance. For most dominant class , the number of data points are around 13444 - Counter(3: 13444, 0: 8057, 4: 3163, 6: 1978, 5: 150, 2: 74, 1: 57). Hence data balancing was necessary.

To overcome imbalance, oversampling the minority class is done using Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al.; 2002). SMOTE first picks a minority class instance, say a, at random and finds its 'k' nearest minority class neighbors. The synthetic instance is then created by choosing one of the k nearest neighbors, say b, at random and connecting a and b to form a line segment in the feature space. The synthetic instances are generated as a convex combination of the two instances picked i.e., a and b. The balanced data for the most dominant class referred above looked like - Counter(1: 13089, 0: 13030, 4: 12959, 3: 12952, 6: 12939, 5: 12901, 2: 12882). The ML models were applied on the balanced data.

Artificial recurrent neural network(RNN) models are employed in the study. Below models are deployed for six class of data picked from feature selection - 'features on hypothesis', 'all features', 'features with high fisher scores', 'features on hypothesis showing normal distribution', 'features with high fisher scores and showing normal distribution' and 'all the features showing normal distribution'. The train and test data are split at 70/30. 'ReLU' activation function is used while encoding and decoding, except for the last decode dense layer, uses 'softmax' which maps the values in the range of [0,1] which add up to 1. All the four models use sequential topology, categorical cross entropy loss function, 'adam' optimizer function and 'minimum' mode. Each of these four models have seven dense layers with 64 units each. 'Accuracy' is used as the evaluation metric.

- **LSTM**: Long Short-Term Memory (LSTM) networks is a kind of RNN model that deals with the vanishing gradient problem. It learns to keep the relevant content of the sentence and forget the non relevant ones based on training. This model preserves gradients over time using dynamic gates that are called memory cells.

  The input layer has 128 units and each of the 7 dense layers have 64 units. 'Softmax' activation function was used. The model was run at 40 epochs and batch size of 64.

- **Autoencoder LSTM**: In the study stacked autoencoder is used. A stacked autoencoder is a neural network consist several layers of sparse autoencoders where output of each hidden layer is connected to the input of the successive hidden layer.

'Adam' optimizer and 'mse' loss function are used to encode the input data into a data of dimension 32. The autoencoder is then trained with 140 epochs, batch size of 64, 'categorical crossentropy' loss function and 'adam' optimizer function.

- **Bi-LSTM**: The Bidirectional LSTM (Bi-LSTM) model keeps two states for forward and backward inputs created by two distinct LSTMs. The first LSTM is a normal sequence that starts at the beginning of the phrase, but the second LSTM feeds the input sequence backwards. The goal of a bi-directional network is to gather data from nearby inputs. It is designed to learn faster than one-directional approach although it depends on the task.

  The input layer has 128 units and each of the 7 dense layers have 64 units. 'Softmax' activation function was used. The model was run at 40 epochs and batch size of 64.

- **Autoencoder Bi-LSTM**: Stacked autoencoder is used in the Bi-LSTM model to perform multi class classification.

  'Adam' optimizer and 'mse' loss function are used to encode the input data into a data of dimension 32. The autoencoder is then trained with 120 epochs, batch size of 64, 'categorical crossentropy' loss function and 'adam' optimizer function.

5 Evaluation

The RNN models were implemented for six sets of features. The overall implementation involved two experiments : employing dropout function and without using dropout function. Accuracy of the models are listed in the tables below.

5.1 RNN Models without Dropout Function

From below table 2 Bi-LSTM model shows better results for all six feature sets with accuracy ranging from 77% to 79%. The highest, 79%, for features on hypothesis showing normal distribution. Results from LSTM models are not far behind. For features on hypothesis the model showed 77% accuracy. Autoencoder models did not show great results. The maximum accuracy obtained from autoencoder LSTM and autoencoder Bi-LSTM are is 54% and 57%

Table 2: Accuracy of Models without Dropout Function

Models/Features	Features on Hypothesis	All Features	Features with High Fisher Scores	Features on Hypothesis Showing Normal Distribution	Features with High Fisher Scores and Showing Normal Distribution	All Features Showing Normal Distribution
LSTM	77%	76%	75%	76%	73%	76%
Bi-LSTM	78%	78%	78%	79%	77%	78%
Autoencoder LSTM	28%	46%	43%	48%	54%	50%
Autoencoder Bi-LSTM	48%	47%	55%	53%	53%	57%

5.2 RNN Models with Dropout Function

Droupout function is a regularization technique used to reduce over fitting in Neural Networks. The dropout values are set to 0.25 and 0.20 on alternative dense layers.

From the below table 3, Bi-LSTM still proves to be the best of the four models with accuracy ranging from 73% to 78%. The highest, 78%, is for set of features showing normal distribution. The best result of the encoder models is 56% for normally distributed features with high Fisher scores.

Table 3: Accuracy of Models with Dropout Function

Models/Features	Features on Hypothesis	All Features	Features with High Fisher Scores	Features on Hypothesis Showing Normal Distribution	Features with High Fisher Scores and Showing Normal Distribution	All Features Showing Normal Distribution
LSTM	72%	73%	68%	74%	72%	73%
Bi-LSTM	76%	76%	73%	77%	75%	78%
Autoencoder LSTM	40%	40%	43%	46%	48%	43%
Autoencoder Bi-LSTM	45%	46%	43%	53%	56%	54%

5.3 Discussion

Overall, the results obtained in the rst experiment is better. While accuracy obtained from LSTM model in the second experiment is less compared to that obtained in the earlier experiment, the accuracy showed by two autoencoder models showed no greater change upon employing the dropout function. The best result obtained from the study is 79%, when the Bi-LSTM model is trained and tested taking into consideration the normal distribution features on hypothesis.

The accuracy is better than the results obtained earlier (see Table 1), from Random Forest, the model with highest accuracy, 73%, obtained in the experiment conducted by Guan et al. (2017), listed in 2.5. With high accuracy, no over tting and class imbalance eliminated, the Bi-LSTM model is certainly well trained and can be deemed the best so far to predict the popularity of news published online.

6 Conclusion and Future Work

The goal of the study is to enhance the accuracy of news popularity forecast.The other objective of the research is to compare the results with that obtained from models employed by other researchers earlier. The amount of likes, comments, and shares may all be used to determine the popularity of a piece of content. People disseminate information that they believe is vital. The variable for predicting popularity in this work is the number of shares of online news items. The data set used in this study is taken from UCI machine learning repository. News in the dataset has various domains like lifestyle, entertainment, business, social media, technology and world. This paper presents LSTM and Bi-LSTM, reinforced neural network models with and without auto-encoder, applied on six combinations of feature subsets. Two experiments are conducted, one with dropout function and another without it.

LSTM model without dropout function used in this research showed up to 79% accuracy, taking into account the features on hypothesis showing normal distribution. There is a signi cant improvement in accuracy compared that of models employed earlier. According to the ndings of the experiments, the suggested method can be used to solve the popularity prediction problem better than before.

With better computational power the models can be trained e ectively, multiple combinations of features can be tried and tested, and di erent network topology can be tested and compared. This is the future work suggested. Also, trying the models for other datasets would give interesting results.

References

Ananda, T. and Sandi, V. (2019). Students' experiences of using vlog to learn english, *Journal of Foreign Languange Teaching and Learning* **4**.

Aragon, P., Gomez, V., Garc a, D. and Kaltenbrunner, A. (2017). Generative models of online discussion threads: state of the art and research challenges, *Journal of Internet Services and Applications* **8**(1).

Aragon, P., Gomez, V. and Kaltenbrunner, A. (2017). To thread or not to thread: The impact of conversation threading on online discussion.

Auxier, B. and Anderson, M. (2021). Social media use in 2021.
URL: *https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/*

Bandari, R., Asur, S. and Huberman, B. A. (2012). The pulse of news in social media: Forecasting popularity, *CoRR* **abs/1202.0332**.
URL: *http://arxiv.org/abs/1202.0332*

Beck, J., Huang, R., Lindner, D., Guo, T., Ce, Z., Helbing, D. and Antulov-Fantulin, N. (2019). Sensing social media signals for cryptocurrency news, *CoRR* **abs/1903.11451**.
URL: *http://arxiv.org/abs/1903.11451*

Carral, D., Dragoste, I., Gonzalez, L., Jacobs, C., Krøtzsch, M. and Urbani, J. (2019). Vlog: A rule engine for knowledge graphs, *in* C. Ghidini, O. Hartig, M. Maleshkova, V. Svatek, I. Cruz, A. Hogan, J. Song, M. Lefrancois and F. Gandon (eds), *The Semantic Web { ISWC 2019*, Springer International Publishing, Cham, pp. 19{35.

Chapman, P., C. J. K. R. K. T. R. T. S. C. and Wirth, R. (2000). Crisp-dm 1.0: Step-by-step data mining guide.
URL: *https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72*

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique, *Journal of Arti cial Intelligence Research* **16**: 321{357.
URL: *http://dx.doi.org/10.1613/jair.953*

Fernandes, K., Chicco, D., Cardoso, J. and Fernandez, J. (2018). Supervised deep learning embeddings for the prediction of cervical cancer diagnosis, *PeerJ Computer Science* **4**: e154.

Fernandes, K., Vinagre, P. and Cortez, P. (2015). A proactive intelligent decision support system for predicting the popularity of online news, *in* F. Pereira, P. Machado, E. Costa and A. Cardoso (eds), *Progress in Arti cial Intelligence*, Springer International Publishing, Cham, pp. 535{546.

Glowacki, M., Narayanan, V., Maynard, S., Hirsch, G., Kollanyi, B., Neudert, L.-M., Howard, P. N., Lederer, T. and Barash, V. (2018). Demtech: News and political information consumption in mexico: Mapping the 2018 mexican presidential election on twitter and facebook.
URL: *https://demtech.oii.ox.ac.uk/research/posts/news-and-political-information-consumption-in-mexico-mapping-the-2018-mexican-presidential-election-on-twitter-and-facebook/*

Guan, X., Peng, Q., Li, Y. and Zhu, Z. (2017). Hierarchical neural network for online news popularity prediction, *2017 Chinese Automation Congress (CAC)*, pp. 3005{3009.

Haghighi, S., Jasemi, M., Hessabi, S. and Zolanvari, A. (2018). Pycm: Multiclass confusion matrix library in python, *Journal of Open Source Software* **3**: 729.

Harrison, X., Donaldson, L., Correa, M., Evans, J., Fisher, D., Goodwin, C., Robinson, B., Hodgson, D. and Inger, R. (2018). A brief introduction to mixed e ects modelling and multi-model inference in ecology, *PeerJ* **6**: e4794.

Housley, W., Webb, H., Williams, M., Procter, R., Edwards, A., Jirotka, M., Burnap, P., Stahl, B., Rana, O. and Williams, M. (2018). Interaction and transformation on social media: The case of twitter campaigns, *Social Media + Society* **4**: 205630511775072.

Kiranmai, S. and Ahuja, L. (2018). Data mining for classi cation of power quality problems using weka and the e ect of attributes on classi cation accuracy, *Protection and Control of Modern Power Systems* **3**.

Litvinenko, A. (2021). Youtube as alternative television in russia: Political videos during the presidential election campaign 2018, *Social Media + Society* **7**: 205630512098445.

Liu, C., Wang, W., Zhang, Y., Dong, Y., He, F. and Wu, C. (2017). Predicting the popularity of online news based on multivariate analysis, *2017 IEEE International Conference on Computer and Information Technology (CIT)*, pp. 9{15.

Ma, C., Yan, Z. and Chen, C. (2017). Larm: A lifetime aware regression model for predicting youtube video popularity, *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* .

Machado, C., Kira, B., Hirsch, G., Marchal, N., Kollanyi, B., Howard, P. N., Lederer, T. and Barash, V. (2015). News and political information consumption in brazil: Mapping the rst round of the 2018 brazilian presidential election on twitter.
URL: *https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/10/machado*$_{et}$$_{al}$$.pdf$

Namous, F., Rodan, A. and Javed, Y. (2018). Online news popularity prediction, *2018 Fifth HCT Information Technology Trends (ITT)*, pp. 180{184.

Nashaat, M. and Miller, J. (2021). Improving news popularity estimation via weak supervision and meta-active learning.

Obiedat, R. (2020). Predicting the popularity of online news using classi cation methods with feature ltering techniques.

Rathord, P., Jain, A. and Agrawal, C. (2019). A comprehensive review on online news popularity prediction using machine learning approach, *SMART MOVES JOURNAL IJOSCIENCE* **5**: 7.

Ren, H. and Yang, Q. (2015). Predicting and evaluating the popularity of online news. **URL:** *https://cs229.stanford.edu/proj2015/328*$_{report.pdf}$

Riganello, F., Larroque, S., Bahri, M., Heine, L., Martial, C., Carriere, M., Charland-Verville, V., Aubinet, C., Vanhaudenhuyse, A., Chatelle, C., Laureys, S. and Di Perri, C. (2018). A heartbeat away from consciousness: Heart rate variability entropy can discriminate disorders of consciousness and is correlated with resting-state fmri brain connectivity of the central autonomic network, *Frontiers in Neuroscience* **12**.

Sitapara, R., Kotian, P., Chaudhary, P. and Kamble, S. (2018). Machine learning methods for online news popularity prediction.
URL: *https://d1wqtxts1xzle7.cloudfront.net/59025447/IC-CSOD*$_2$01820190425 $-$ 2462 $-$ $m2c0gy - with - cover - page - v2.pdf?Expires$ = 1629051568$Signature$ = $Q19pujQ4DS$ $etHUS1NdF$ $-$ $Zr3CoXxA5aW2Y05bxHACYyVkYyIM$ $-$ $x5JEuyKcBm1zevanJAg6ZkO5RDVTxM73tHLl5r40yz87HFuHg4oTCTHNkiENIa$ $pqgNiuQL9oMqwVa7OZN1F2JuENsRNblxxdbrxERlUUqjyfCux6gcg7LCa9odT9rZw$ $EuPTN23tHxCCeUaBCHL1cVEMT2VXCQjJg2C1UcvKpmRJzCwELQhBO1Ca1$ $Xq2D\ hA4 - tXC\ 2DqTS6okwzMbWAs2VVpL0qzLuDlp6MD -$ 71 $8FEu8u-POum88N8S\ 2gEAyWf9-PzsOStd0of0TLu-\ O0ptnkJCqwxOqf1ukA\_\_$ $Key - Pair - Id$ = $APKAJLOHF5GGSLRBV4ZApage$ = 328

Tatar, A., Antoniadis, P., de Amorim, M. D. and Fdida, S. (2012). Ranking news articles based on popularity prediction, *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 106{110.

Tatar, A., Dias de Amorim, M., Fdida, S. and Antoniadis, P. (2014). A survey on predicting the popularity of web content, *Journal of Internet Services and Applications* **5**.

Uddin, M. T., Patwary, M. J. A., Ahsan, T. and Alam, M. S. (2016). Predicting the popularity of online news from content metadata, *2016 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pp. 1{5.

Wicaksono, A. S. and Supianto, A. A. (2018). Hyper parameter optimization using genetic algorithm on machine learning methods for online news popularity prediction, *International Journal of Advanced Computer Science and Applications* **9**.

Xiong, J., Yu, L., Zhang, D. and Leng, Y. (2021). Dncp: An attention-based deep learning approach enhanced with attractiveness and timeliness of news for online news click prediction, *Information Management* **58**(2): 103428.
URL: *https://www.sciencedirect.com/science/article/pii/S0378720621000021*