Machine Learning for Predicting Stock Market Movement using News Headlines

Yang Liu*, Jelena Trajkovic*, Hen-Geul (Henry) Yeh†, Wenlu Zhang*
*Department of Computer Engineering and Computer Science, California State University Long Beach, CA, 90840
†Department of Electrical Engineering, California State University Long Beach, CA, 90840

***Abstract*—There are many factors that affect performance of stock market, such as global and local economy, political events, supply and demand, and out of the ordinary events, as COVID-19 pandemic. The factors may not only influence the stock market movement, but also influence each other. We propose to observe the movement of Dow Jones Industrial Average in relations to daily news. We use top-5 news headlines from Reddit to create 1-Day and 5-Day models to predict if Dow Jones Industrial Average movement will be in Down and Up direction from the moment the market opens till it closes. We propose use of shallow (traditional) Machine Learning algorithms and Deep Learning algorithms. Additionally, we explore the effect of word representation, using TF-IDF and GloVE approaches. Moreover, we evaluate our models in terms of accuracy of prediction on data sets containing data before pandemic and during pandemic. Our models show that Deep Learning models uniformly have higher accuracy than Machine Learning ones. Convolution Neural Network with TF-IDF and 5 Days prediction performs the best for the dataset before the pandemic with accuracy of 59.6%. Gated Recurrent Unit (GRU), a class of Recurrent Neural Networks, with GloVe and 1 Day prediction outperforms the other models for dataset during the pandemic with the accuracy of 62.9%.**

***Index Terms*—Stock Movement, Text Classification, Text Mining, Machine Learning, Deep Learning**

I. Introduction and Motivation

Analyzing stock market performance and using the analysis for short and long term predication of stock market movement is an important and difficult problem. Stock market movement can be easily affected by political strategy, economy stability, trade war, unemployment rate, the expectations of investors and some unprecedented event such as the recent pandemic, COVID-19, etc. Therefore, it is always essential and challenging for the stockholder to be capable of accurately predicting the stock values. The magnitude of change of each influencing factor, and the nature of change (positive or negative change) create an intricate dependency between the factor and the stock market movement. This research area is broad and includes risk assessment and portfolio management, but we will focus on prediction of Dow Jones Industrial Average (or Dow Index) that tracks and combines performance on stock markets for 30 large, publicly-owned companies trading on the US stock market.

We will adopt Dow index as an indicator of stock market movement. To illustrate the stock movement we plot distribution of weekly, monthly and annual change of Dow Index in Figures 1, 2, and 3, respectively. The aggregate data for the figures are scraped from [1] during years 2008 - 2020 and the

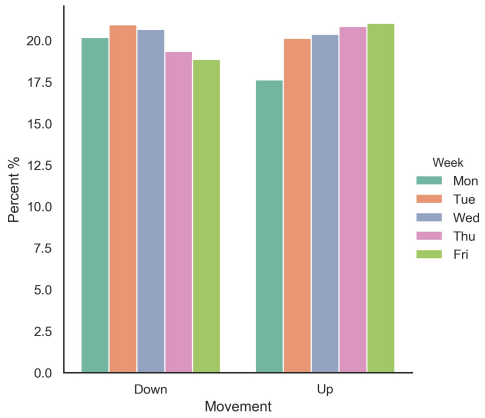

Fig. 1. Weekly Distribution of Up and Down Movement

difference between opening and closing price was computed to indicate movement Down or Up. The Y-axes in Figure 1 shows percentage of days, with respect to total number of days in 2008 - 2020., when the stock market went down or up, during each day of the week. Similarly, in Figure 2, and 3, each bar represents cumulative percentage of occurrences within a month and within each year (2008, 2009, ...), respectively. We can observe that movement is distributed pretty evenly across all data instances of the inputs, and the complex movement patterns are not easily noticeable in these visualizations.

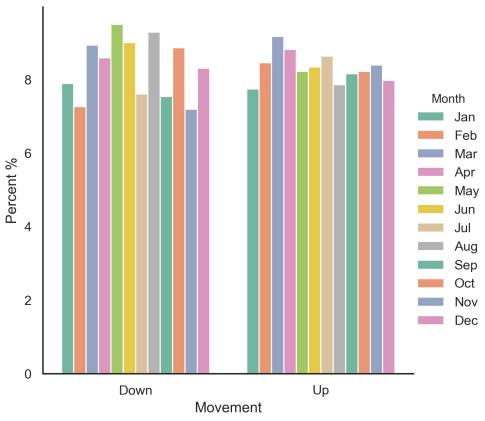

Fig. 2. Monthly Distribution of Up/Down Movement

978-1-7281-8744-0/20/$31.00 ©2020 IEEE

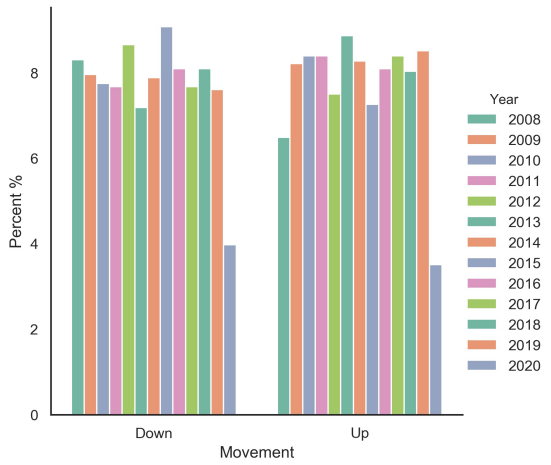

Fig. 3. Yearly Distribution of Up and Down Movement

Fig. 4. News Words in Down Movement

Additionally, we propose to use the headlines to predict the movement of the Dow index. In this study, we will use top-5 highest rated news headlines from Reddit [2]. Figures 5 and 4 visualize frequent words from the headlines in word-cloud. As seen in the figures, the most influential words for the Down and Up movement are very similar.

In this paper, we will be focusing on comparing the shallow Machine Learning (ML) and Deep Leaning (DL) approaches to predict the Dow Index movement based on the news headlines. Additionally, we will explore and compare use of two different text representation methods as an addition to all the approaches. Finally, we will consider two data sets: one that includes dates before the COVID-19 pandemic and the other that includes the dates during the COVID-19 pandemic up to the June 17th 2020 date, to study how our models handle unprecedented events due to pandemic.

II. Related Work

Stock movement analysis is a pertinent and challenging topic and there are a wide variety of studies that approached it from different perspectives. Therefore we present the related work in the several categories, both in terms of the topic and techniques that relate to our research.

Fig. 5. News Words in Up Movement

A. Stock Market Prediction

Nabipour et al. [3] implements six tree-based and three neural networks based algorithms to generate models to predict four different market groups. The models were used for short-medium- and long-term prediction (1 - 30 days). While the method has similar goal, it bases the prediction on history of markets itself, without explicitly including the external factors that might be very influential. Wang et al. [4] classifies the financial time series prediction into two major classes: the statistical models and machine learning approaches, where we focus only on the latter. Ballings at al. [5] performed experimental evaluation of ensemble methods (Random Forest, AdaBoost and Kernel Factory) and single classifier models (Neural Networks, Logistic Regression, Support Vector Machines and K-Nearest Neighbor), and found that Random Forest outperforms the other models. The scope of this study differs from ours as it classifies the stock price growth within buckets of predetermined % amount one year ahead.

Huang et al. [6] uses wavelet to build the Support Vector Machine (SVR) kernel to forecast financial time series using Nasdaq composite index. They compared it to several kernel implementations and showed that the biorthogonal spline wavelet Bior(4,4) kernel SVR has better accuracy compared to standard kernel functions. Liang et al. [7] uses nonlinear SVR to build the model where web time series information is mapped to stock price time series, and showed that the use of web information was effective in improving forecasting accuracy when compared with single SVR. In this work, we will evaluate SVM as one of the proposed ML methods.

B. Stock Prediction and News Headlines

A subset of the stock prediction work uses content of news to predict stock movement. Currie et al. [8] presents an open-source project that expands on the work of [9] that proposes methods for sentiment analysis of short texts. Currie et al. [8] uses GloVe's [10] applied on Convolution Neural Network (CNN) and Recurrent Neural Network (RNN). About 300 various project, including the above mentioned one, work on the same pre-pandemic dataset (2008-2016) and [10] reports accuracy in range of 50% to 54%. Our project differs significantly as our experiments considers prediction during the

pandemic, and our experimental results show accuracy up to 62.9%.

C. TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) [11] is a numerical statistical tool for obtaining a matrix representation from converted text documents and commonly used in Natural Language Processing (NLP) related applications such as information retrieval and text mining, etc. TF-IDF improves the basic Bag of Words (BOW) [12] approach for converting documents to vectors by accounting for the relevance of words to a particular document compared to other documents in the corpus. TF-IDF score of a word is the product of two statistics terms. The first one is term frequency, which accounts for the relevance of the word to a document. The second term is inverse document frequency, which accounts for how common a word is in the corpus. Thus, words that are common in every document will get low scores. The output of TF-IDF is a sparse vector with high dimension for each document where the total number of non-zero elements is equal to the number of unique words in a document.

D. GloVe

TF-IDF representation does not consider the context information of words. Word embedding is another popular representation approach which is capable of capturing the context of a word in a document. In particular, word embedding models map words to low dimensional vectors, which can capture the semantic relationships between words. Word2Vec [13] is the first word embedding model that use shallow neural network to learn word embeddings. The Global Vectors for Word Representation (GloVe) [14] was proposed after Word2Vec (a basic word embedding approach), and has become one of the most popular pre-trained databases of word embeddings. GloVe is a log-bilinear regression algorithm to generate vector representations of words. Different from Word2Vec, GloVe model utilizes the word co-occurrences matrix to capture the global statistics, which increases the computational efficiency and improves the performance on uncommon words in the corpus.

III. Proposed Technique

Analyzing stock market performance has been a popular research topic in finance and marketing areas. In this project, we aim to find out the effect of global/national big events happened in the past several hours or days to the stock market movements. The proposed prediction modeling technique to predict stock market trends based on the news headlines is outlined in this section.

A. Data Preparation

We utilize the daily world news headlines from Reddit via Pushshift API [15] to predict the movement of Dow Jones Industrial Average such as “Up” and “Down”, where the “Up” movement means that closing value of Dow Jones Industrial Average is larger than the opening values of the Dow Jones Industrial Average in the same day, and the “Down” movement means that closing value of the Dow Jones Industrial Average is less than the opening values of the Dow Jones Industrial Average in the same day. Specifically, we collect the news headlines for over 12 years. According to the volumes of “likes” and “dislikes” for a single news headline, we rank and pick top 5 news headlines for each day. Three examples of news headlines published before 2020 and the corresponding label are shown in Table I. We also want to explore if the unprecedented event, such as COVID-19, will affect the economy stability, therefore another three examples of news headlines posted between March and June in 2020 are provided in Table II.

B. Data Preprocessing and Word Representation

As our input data is in the text format, we apply Natural Language Processing techniques that clean raw text in order to improve model performance. In this work, we use the following preprocessing techniques:

- converting every word to lower case;
- removing all non-ASCII characters;
- removing ”stop words” (such as “the”, “a”, “an”, “in”) using NLTK library [16];
- removing of punctuation;
- implementing lemmatization that groups the different inflected forms of a word so that they can be analyzed as a single word (i.e. “rocks” will be grouped with “rock”; “better” will be grouped with “good”; “am”, “are”, “is” will be grouped with “be”).

The result of the preprocessing is in text format. We propose to explore two word representation techniques: TF-IDF and GloVe, described in II-C and II-D, respectively. While the former transforms the input text into sparse vector, the latter uses matrix format. Both techniques not only represent the text in numerical values, but aim to extract unique word, global statistics and relationships between the words in the text. The resulting dataset is the input for all Machine Learning and Deep Learning algorithms that we implement and evaluate in this work.

C. Machine Learning Algorithms

a) Shallow Machine Learning Algorithms: Logistic Regression is a commonly used Supervised Machine Learning algorithm, usually represents probability for the binary classification. Support Vector Machine (SVM) is another popular classification model. The intuition of SVM is to find the optimal hyperplane that separate the dataset [17]. SVM is robust to handle nonlinear separable dataset by using kernel. In this study, we use RBF kernel to classify high-dimensional stock movement dataset. Random Forest (RF) is an ensemble technique that consists of a large collection of independent decision trees. One advantage of RF is that it can average noises, and eliminate the variance among different decision trees. In our implementation of RF, we set the number of trees as 100, minimum samples per split are assigned to 2.

b) Convolutional Neural Network: In recent years, Deep Learning has made extraordinary improvements in vision [18]–[20] and speech recognition [21]. Convolutional Neural Network (CNN) is one of the most popular deep learning models. CNN algorithm was originally designed for computer vision problems since it can utilize the convolution and pooling operation to preserve the spatial information. Convolutional Neural Network has also shown great promise in semantic analysis [22], language modeling [23], sentence classification [24] and other Natural Language Processing applications [25]. In this project, we got the insight from VGG [26] models to build 5 convolution layers with the kernel size as 3. Furthermore, our proposed CNN models are all based on word representations, such as TF-IDF or GloVe.

c) Recurrent Neural Network: Recurrent Neural Networks (RNNs) [27]–[30] have been widely used to recognize patterns in sequences of data such as time-series data, text data, and biological data which contain sequential dependencies among the features. Different from CNNs and other deep neural networks, RNNs add a temporal dimension to take into account the sequential order of data (e.g., words in a sentence). The position in the sequence is also referred as its time-stamp. At each time step, RNN uses recurrent computation (recurrent layer) to calculate the hidden state of the sequence, which preserves the historical information of the sequence up to the current time step. The hidden state is then used to generate the output at that time step. Because different time steps share the parameters, so the number of RNN model parameters does not grow as the number of time steps increases. Vanilla RNNs suffer from vanish or divergent gradients issues. In practice, Long Short Term Memory (LSTM) [31] and Gated Recurrent Unit (GRU) [32] are two popular RNNs that are widely used for many NLP applications. In LSTM, each unit contains input gate, output gate and forget gate. Cell state, hidden state, and input are utilized through these gates. Forget gate decides what previous information is to be thrown away, then the input gate decides what information is to be stored. In the meantime, the three gates interact with each other to update cell states, and hidden states. Gated Recurrent Unit (GRU), it is a simplified version of LSTM, it only uses update gate and reset gate interactively to update hidden states. The reset gate decides how much information to take from previous information, then the update gate decides the strength of contribution of lower layers. However, in unidirectional network, the outputs are mostly based on previous information. Bidirectional RNN [33] aims to capture both of previous and future information, where the input is presented forward and backward ways to two separate recurrent neural networks.

IV. Experimental Studies

A. Experimental Setup

The entire dataset is created using the data from February 7th, 2008 to June 17th, 2020, totaling 3112 data points that were divided into two subsets: 1) subset "Before Pandemic": Feb 7th 2008 to Dec 31st 2019 contains 2996 samples and 2) subset "During Pandemic": Jan 1st 2020 to Jun 17th 2020 contains 116 samples. Tables I and II present examples from each of the subsets. For each row in the tables, we show the Date, Headline, True and Predicted Label. The True Label shows if the Dow Index moved Up or Down at the closing times of trading hours. We show one of the Headlines for each selected data in the Table I. We used GRU models with GloVe embedding to generate the Predicted Label, as it is the most accurate model per Sec IV-B. For the Table I and Table II, "1 Day prediction" was used. The proposed Deep Learning method is capable of capturing enough context features to get the correct classification label for Dow Jones Industrial Average. Our model can also correctly predict the stock market movement due to some unrepresented event such as COVID-19 pandemic.

TABLE I
News headlines examples Before Pandemic

Before Pandemic Sample			
Date	Headline	True Label	Predicted Label
2016-02-24	PSA Peugeot-Citroen, ninth largest car manufacturer in the world made a net profit of 989 million dollar in 2015. It gave its employees a Euros 2,000.00 pays bonus but no dividends to its shareholders.	Up	Up
2017-08-22	"Cyborg" bacteria deliver green fuel source from sunlight - Scientists have created bacteria covered in tiny semiconductors that generate a potential fuel source from sunlight, carbon dioxide and water.	Up	Up
2018-09-25	World leaders literally laugh at Trump when he claims he has accomplished more than almost any administration	Down	Down

There are 6 experiments in our study, that were performed as follows:

- To evaluate our models "Before Pandemic" we use only the "Before Pandemic" data subset and explore use of TF-IDF and GloVe (Tables III and IV, respectively). Additionally, we explore how the duration of time we use to make the prediction affects our results. We use same day's headlines and 5 days' headlines (columns "1 Day" and "5 Days", respectively) to predict if the Dow Index moved Up or Down from the opening to the closing of the market. The split size of training data and validation data is 0.2.
- To evaluate our models "During Pandemic" we re-train our models using "Before Pandemic" data subset for training, and validate the models using "During Pandemic" data subset for validation. We also evaluate both TF-IDF and GloVe, but only utilize the same day's headlines for prediction, due to the relatively small number of data points in "During Pandemic" data subset.

TABLE II
NEWS HEADLINES EXAMPLES DURING PANDEMIC

During Pandemic Sample			
Date	Headline	True Label	Predicted Label
2020-03-13	Trump says Google building coronavirus test website, Trump says	Up	Up
2020-03-24	Coronavirus: Uganda receives Jack Ma consignment.The overall shipment includes over 1.5 million laboratory diagnostic test kits and over 100 tonnes of infection prevention and control commodities	Up	Up
2020-06-09	Coronavirus: This is not the last pandemic - In the last 20 years, we have had six significant threats - SARS, MERS, Ebola, avian influenza and swine flu. We dodged five bullets but the sixth got us.	Down	Up

For model generation we apply shallow ML methods: Logistic Regression, SVM, and Random Forest; and Deep Learning (DL) methods: CNN, LSTM, and GRU. We compare the accuracy of ML and DL methods.

To build CNN models, we apply 5 convolutional layers with different number of kernels (16, 32, 64, 128, 256), and use ReLU activivation after each convolution operation. All convolutional layers are using filter with size 3, followed by max pooling with size 2. Finally, we apply global maxpooling layer. We optimize CNN model using RMSprop optimizer (learning rate=0.001, rho=0.9, epsilon=1e-07). Loss function used is a binary cross entropy loss function.

To build both RNN (LSTM and GRU) models, we use 2 layers of bidirectional RNN model followed by a Sigmoid activation layer. We set the hyper-parameters as follows: there are 32 units in first layer, and 5 units in second layer; recurrent dropout is 0.5. We optimize RNN models using RMSprop optimizer (learning rate=0.001, rho=0.9, epsilon=1e-07). Loss function used is a binary cross entropy loss function.

We compare performance across different models using the validation accuracy that is computed as per Equation 1.

$$\text{Accuracy} = \frac{\text{Correctly Predicted}}{\text{Correctly Predicted + Incorrectly Predicted}} \quad (1)$$

B. Experimental Results and Discussion

In Table III and Table IV, we utilize all the news headlines from February 7th, 2008 to December 31st, 2019 to evaluate the efficiency of the several proposed Machine Learning and Deep Learning methods. The performance of all the models using TF-IDF are summarized in Table III and the experimental results of the same models using GloVe are listed in Table IV. In general, Deep Learning algorithms including LSTM, GRU and CNN outperform Machine Learning methods

TABLE III
EXPERIMENTAL RESULTS USING TF-IDF VECTOR FOR 1 DAY & 5 DAYS PREDICTION BEFORE PANDEMIC

Method	Accuracy	
	1 Day	5 Days
Logistic Regression	0.470	0.503
SVM	0.523	0.509
Random Forest	0.492	0.502
CNN	0.573	0.596
LSTM	0.573	0.556
GRU	0.555	0.576

TABLE IV
EXPERIMENTAL RESULTS FOR GLOVE EMBEDDING FOR 1 DAY & 5 DAYS PREDICTION BEFORE PANDEMIC

Method	Accuracy	
	1 Day	5 Days
Logistic Regression	0.482	0.510
SVM	0.543	0.510
Random Forest	0.512	0.504
CNN	0.505	0.510
LSTM	0.515	0.543
GRU	0.565	0.536

in both word representations. We observe that Convolution Neural Network with five consecutive days' data achieves the best result if choosing TF-IDF vectorization. Nevertheless, due to the limited number of news headlines, if we choose GloVe embedding as word representation, the strength of Deep Learning methods is limited by preserving sequence-to-sequence neighboring context.

Table V summarizes the evaluation results of the news headlines during novel Coronavirus pandemic. We plan to find out the correlation between unrepresented event such as COVID-19 and stock market movement. Due to the limited number of news headlines, we decide to implement our models only based on the 1 Day perdition. GRU outperforms the other models. We can prove that by using the time series data, Recurrent Neural Networks and Convolutional Neural Networks can capture more dependency features than Logistic Regression, Support Vector Machine and Random Forest.

V. CONCLUSION AND FUTURE WORK

In this work, we analyze the stock market activity performance using daily world news headlines from Reddit. In particular, we propose novel Convolutional Neural Network and Recurrent Neural Network architectures to find out the effect of global big events to the stock market movements. We implement our model on the movement of Dow Jones Industrial Average from February 2008 to June 2020. We also provide several news headlines examples to show the efficiency and effectiveness of our proposed Deep Learning model. Especially, we prove that our model is capable of predicting the economy stability due to the unrepresented events such as COVID-19. In the future, we plan to employ Word2Vec word embedding and deep transformers to improve

TABLE V
EXPERIMENTAL RESULTS USING TF-IDF VECTOR & GLOVE EMBEDDING FOR 1 DAY PREDICTION DURING PANDEMIC

Method	Accuracy	
	TF-IDF	GloVe
Logistic Regression	0.534	0.569
SVM	0.483	0.595
Random Forest	0.466	0.526
CNN	0.509	0.379
LSTM	0.560	0.595
GRU	0.603	0.629

the performance of stock market prediction using financial related headline news.

ACKNOWLEDGMENT

REFERENCES

[1] "Yahoo finance," https://finance.yahoo.com/.

[2] "Reddit news," https://www.reddit.com/r/news.

[3] M. Nabipour, P. Nayyeri, H. Jabani, and A. Mosavi, "Deep learning for stock market prediction," *arXiv preprint arXiv:2004.01497*, 2020.

[4] J.-Z. Wang, J.-J. Wang, Z.-G. Zhang, and S.-P. Guo, "Forecasting stock indices with back propagation neural network," *Expert Systems with Applications*, vol. 38, no. 11, pp. 14 346–14 355, 2011.

[5] M. Ballings, D. Van den Poel, N. Hespeels, and R. Gryp, "Evaluating multiple classifiers for stock price direction prediction," *Expert Syst. Appl.*, vol. 42, no. 20, p. 7046–7056, Nov. 2015. [Online]. Available: https://doi.org/10.1016/j.eswa.2015.05.013

[6] C. Huang, L. Huang, and T. Han, "Financial time series forecasting based on wavelet kernel support vector machine," in *2012 8th International Conference on Natural Computation*, 2012, pp. 79–83.

[7] X. Liang, R.-C. Chen, Y. He, and Y. Chen, "Associating stock prices with web financial information time series based on support vector regression," *Neurocomputing*, vol. 115, p. 142–149, 09 2013.

[8] D. Currie, "Predicting-the-dow-jones-with-headlines," https://github.com/Currie32/Predicting-the-Dow-Jones-with-Headlines, 2017.

[9] X. Wang, W. Jiang, and Z. Luo, "Combination of convolutional and recurrent neural network for sentiment analysis of short texts," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 2428–2437. [Online]. Available: https://www.aclweb.org/anthology/C16-1229

[10] "Kaggle dataset: Daily news for stock market prediction," url:https://www.kaggle.com/aaron7sun/stocknews.

[11] A. Aizawa, "An information-theoretic perspective of tf–idf measures," *Information Processing & Management*, vol. 39, no. 1, pp. 45–65, 2003.

[12] Z. Harris, "Distributional structure," 1954.

[13] Y. Goldberg and O. Levy, "word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[14] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[15] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The pushshift reddit dataset," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 830–839.

[16] "Nltk library," https://www.nltk.org/book/ch02.html.

[17] H. Trevor, T. Robert, and F. Jerome, "The elements of statistical learning," *Springer*, pp. 119–127, 2008.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[20] S. Christian, L. Wei, J. Yangqing, S. Pierre, R. Scott, A. Dragomir, E. Dumitru, V. Vincent, and V. Vincent, "Going deeper with convolutions," *arXiv:1409.4842v1 [cs.CV]*, 2014.

[21] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2011.

[22] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "Learning semantic representations using convolutional neural networks for web search," in *Proceedings of the 23rd international conference on world wide web*, 2014, pp. 373–374.

[23] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*, 2017, pp. 933–941.

[24] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, p. 1746–1751.

[25] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[26] S. Karen and Z. Andrew, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556v6 [cs.CV]*, 2015.

[27] J. F. Kolen and S. C. Kremer, *A field guide to dynamical recurrent networks*. John Wiley & Sons, 2001.

[28] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *International conference on machine learning*, 2013, pp. 1310–1318.

[29] R. Pascanu, T. Mikolov, and Y. Bengio, "Understanding the exploding gradient problem," *CoRR, abs/1211.5063*, vol. 2, p. 417, 2012.

[30] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and understanding recurrent networks," *arXiv preprint arXiv:1506.02078*, 2015.

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[32] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[33] S. Mangal, P. Joshi, and R. Modak, "Lstm vs. gru vs. bidirectional rnn for script generation," *arXiv preprint arXiv:1908.04332*, 2019.