Natural Language Contents Evaluation System for Multi-class News Categorization Using Machine Learning and Transformers

Duván A. Marrugo, Juan Carlos Martinez-Santos, and Edwin Puertas(✉)

Universidad Tecnológica de Bolívar, Cartagena de Indias, Colombia
{marrugod,jcmartinezs,epuerta}@utb.edu.co

Abstract. The exponential growth of digital documents has come with rapid progress in text classification techniques in recent years. This paper provides text classification models, which analyze various steps of news classification, where some algorithmic approaches for machine learning, such as Logistic Regression, Support Vector Machine, and Random Forest, are implemented. In turn, the uses of Transformers as classification models for the solution of the same problem, proposing BERT and DistilBERT as possible solutions to compare for the automatic classification of news containing articles belonging to four categories (World, Sports, Business, and Science/Technology). We obtained the highest accuracy on the machine learning side, with 88% using Support Vector Machine with Word2Vec. However, using Transformer DistilBERT, we got an efficient model in terms of performance and 91.7% accuracy for classifying news.

Keywords: Text Classification · Automatic Classification · News Classification · Transformer · Machine Learning · Deep Learning

1 Introduction

There is a large amount of data stored in electronic format. With this data, the need has arisen for similar means that can interpret and parse similar data and extract valuable data, which we can use to assist decision-making [2]. Furthermore, we use information digging to remove concealed data from big data sets, an exceptionally integral asset utilized for this reason. According to a report by the consulting firm IDC, in 2025, the volume of data will reach 175 zettabytes, which means the equivalent of 175 times the information generated in 2011 [24].

News information was readily and rapidly available in the last decade. As a result, news is now easily affordable through content providers such as online news services. As mentioned in Ofcom's report on news consumption in 2020 [6], 65% of adults use the Internet as a news platform, compared to 41% in 2016, indicating a significant increase in the availability and growing popularity of online news [5].

J. C. Figueroa-García et al. (Eds.): WEA 2023, CCIS 1928, pp. 115–126, 2023.
https://doi.org/10.1007/978-3-031-46739-4_11

Classifying news text automatically assigns a text news document to a defined class of news items from predefined categories. Different approaches to machine learning, such as artificial neural networks, support vector machines, and decision techniques, can be used to solve the text classification problem [25]. Researchers have approached text classification using various clustering methods, Naive Bayes classifier [3], support vector machines with word2vec, and TF-IDT approaches [15]. There have also been several novel approaches to artificial neural networks. As seen in [14,20,21], neural has fruitfully carried out sentiment analysis.

Approaches are using Recurrent Neural Networks (RNN) to solve NLP problems. The success of RNNs is due to the Long Short Term Memory (LSTM) [7]. Moreover, the incredible versatility of these networks makes it possible to resolve a diversity of problems [13]. Authors in [23] introduce a novel model architecture, Transformer, to counter these two limitations. Furthermore, their proposed technique discards the recurrent architecture to depend solely on the attentional mechanism [12].

This paper presents the solution of new classifiers from a Machine Learning perspective, implementing after preprocessing, feature extraction, and three supervised classification methods for developing news classification models. Furthermore, Transformer models Bert and DistilBert will use the pre-trained language representation. Training data will be used for both, choosing the one with the best performance in terms of accuracy. This work is as follows: First, it overviews the related work on text categorization. Next, it describes the methodology implemented for classification models. Then, it presents experimental validation. Finally, it summarizes conclusions and future work.

2 Related Work

Nowadays, several studies focus on solving the categorization problem of news articles in different languages using machine learning methods. For example, in [4], in real-time, the authors evaluated the performance of machine learning-based methods for English news from the BBC website to classify them into five topics: business, entertainment, politics, sports, and technology. The classifiers selected for the analysis were Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF). In addition, they used TF-IDF for feature extraction, with the LR method obtaining the highest accuracy of 95.5%.

As before, in [8], the authors evaluated the performance of the BBC corpus news. They used NB, Multilayer Perceptron Neural Network, DT, and RF classifiers. The same feature extraction method, TF-IDF, was implemented. The best-performing way was NB, with 96.8% accuracy. Next, four machine learning methods, SVM, NB, k-Nearest Neighbors (kNN), and Convolutional Neural Network in [9], were applied to analyze text representation models using feature extraction, bag-of-words, and n-gram methods. Using the SVM model with bag-of-words for the 20 Newsgroups and AG's News corpora, an accuracy of 90.8% and 85.14%, respectively, was obtained.

In [16], authors used three text corpora. The first one includes Women, literature, sports, and campus news. The second corpus includes sports, constellation, games, and entertainment news. They obtained as F1 results for corpus 1 and 2 the best results through SVM: 0.86 and 0.71, respectively. For the last corpus, the best estimation was through LR, with an F1 of 0.63.

In [19], authors developed models using ten machine learning methods and pondered the text employing TF-IDF to classify news articles with the Middle East, technology, sports, and business topics. First, they obtained SVM as the best-performing method, with F1 = 97.9. Then, based on the corpora AR-5, KH-7, AB-7, and RT-40, whose names correspond to the number of topics, nine neural network models were implemented [11].

Nowadays, the use of pre-entrant linguistic models is growing, as is the particular case of BERT (Bidirectional Encoder Representations from Transformers) [10], which is like DistriBERT [22]. These contextual representation models, useful for text classification tasks, provide a contextual representation of words different from word-based embedding models, such as word2vec [17] and GloVe [18], where a unique word embedding is produced for each word, regardless of the context.

3 Methodology

3.1 Dataset Description

To solve the multiclass text classification problem, a text corpus from a competition proposed by the National Yang-Ming Chiao Tung University (NYCU) of Kaggle [1], which linked 58 participants, was used. Given the title and content of the news item, we train a model using Transformers to correctly classify the news item into four different categories: World, Sports, Business, and Science/Technology, which records information on 2000 training samples and 400 test samples. The baseline required for the use of a Transformer was 15%. In turn, the dataset provides four columns of information: ID, Category, Title, and Description. Therefore, the data could be more balanced. For example, sports and world news represent 27% of the dataset, while business represents 25% and sci-tech news 21%.

3.2 Classification Process

Figure 1 visually represents the classification process. It begins with reading the dataset stored in CSV files and organizing it into a data frame with four corresponding columns, as explained in Subsect. 3.1. Subsequently, data preprocessing takes place. We conducted an exploratory analysis to identify keywords, acronyms, and abbreviations facilitated by n-gram analysis for feature selection. We analyzed this using Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and Word2Vec methods. These approaches offer advantages such as resource and time efficiency and enhanced prediction accuracy for the model.

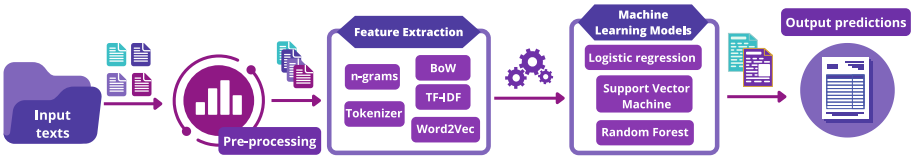


Fig. 1. News classification Machine Learning framework.

3.3 Pre-processing

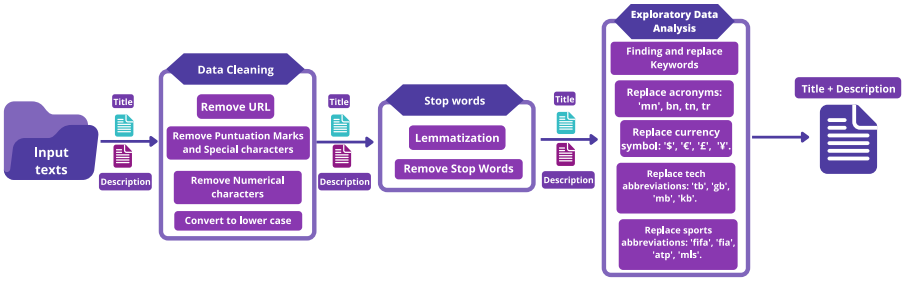


Fig. 2. Data cleaning pipeline.

Data Cleaning. Preprocessing operations consist of cleaning and normalizing the input data to improve the results of the feature extraction stage, i.e., regardless of whether they are manually constructed or automatically learned (in deep learning methods). As shown in the Data Cleaning section in Fig. 2, we focus on eliminating four essential issues: URLs, numeric characters, lowercase conversion, and punctuation marks.

Exploratory Data Analysis. To further normalize the data and reduce the feature space, we implemented additional procedures such as converting all text to lowercase and normalizing slang words and abbreviations. To assess the relationships between words within each news item, we conducted two types of analyses:

– In the Univariate Analysis, we identified words like ‘bn’ (converted to ‘billion’), ‘snday’ (normalized to ‘sunday’), ‘qot’, and ‘bsness’ (expanded to ‘business’). These words were either replaced with their expanded meanings or removed as necessary. Interestingly, this step led to a slight increase in model accuracy, approximately 0.01%, compared to not performing it.

– In the Bivariate Analysis, we employed bigram and trigram analysis to examine word relationships. This analysis unveiled modifications in certain words, such as currency symbols (e.g., "$" to 'dollar', '€' to 'euro') and abbreviations (e.g., 'tb' to 'terabyte', 'gb' to 'gigabyte'). We also observed transformations like 'mn' to 'million', 'bn' to 'billion', and 'tn' to 'trillion'. These normalization techniques contributed to refining the data and enhancing its accuracy.

3.4 Feature Extraction

Our feature extraction approach employed three distinct methods: Bag-of-Words (BoW), TF-IDF (Term Frequency-Inverse Document Frequency), and Word2Vec. For Bag-of-n-grams and TF-IDF, we focused on capturing consecutive word sequences of length "n", denoting 2 and 3 as bigrams and trigrams, respectively. The parameter settings encompassed a frequency range from 1 to 500, an n-gram range spanning (1, 3), and the 'word' analyzer. In the case of Word2Vec, we opted for the Skip-gram model. We conducted meticulous parameter tuning, encompassing factors like the number of features, context window size, and negative sampling, among others, to ensure optimal performance in alignment with prior research findings.

3.5 Classification Using Supervised Algorithms

Initially, we conducted random oversampling with careful control over randomness, setting a specific random state (Random State) to 42. Following this, we implemented cross-validation through random permutation (Shuffle Split) to create a training dataset, accounting for 80% of the total data, and a test dataset, representing the remaining 20%.

We chose three classifiers for our news classification task: Logistic Regression (LR), Random Forest, and Support Vector Machine (SVM) with an RBF kernel, based on their demonstrated effectiveness in previous research. We fine-tuned LR using the 'lbfgs' optimizer with an inverse regularization strength of 1.0, allowing a maximum of 600 iterations for convergence. Random Forest, an ensemble method, incorporated 200 decision tree classifiers, each with a maximum depth of 200, to improve prediction accuracy and prevent overfitting. In the case of SVM, we utilized the 'rbf' kernel for multi-dimensional operations, applied an 'l2' penalty to avoid sparse coefficient vectors, and set the random state to 0 for consistency. These classifiers, each with its specific configuration, were applied to address our news classification task effectively.

3.6 Transformers

Architecture: The critical component of this architecture is the self-attention layer (A), which intuitively allows the encoder to look at other words in the

input sentence whenever processing one of its words. Stacking multiple layers of this type creates a multi-head attention (MHA) layer, as shown in Fig. 3. Then, we condensed the individual outputs into a single matrix by concatenating the head outputs and passing the result through a linear layer.

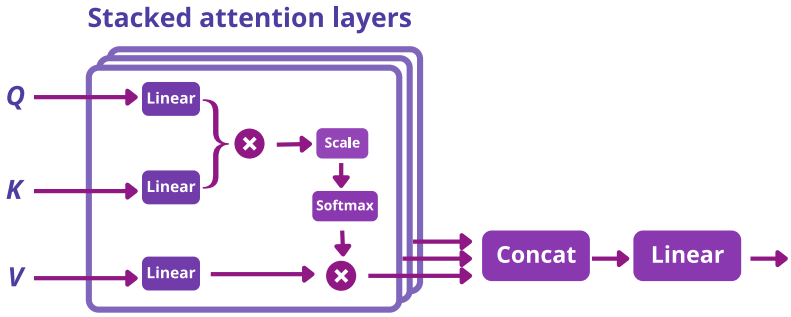


Fig. 3. The multi-head attention layer used in the Transformer architecture.

– **Encoder:** In the encoding part, the input embeddings are multiplied by three separate weight matrices, as indicated in Eq. 1, Q (queries), K (keys), and V (values), to generate different word representations.

$$\begin{aligned} Q &= X \cdot W_Q \\ K &= X \cdot W_K \\ V &= X \cdot W_V \end{aligned} \tag{1}$$

W_Q, W_K, $W_V \in \Re^{dim \times d_k}$ are the learned weight matrices. Eventually, we obtain the representation of each word by multiplying the scaled term with the V matrix containing the input representation. We define this operation in Eq. 2.

$$Z = A(Q, K, V) = S\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) \cdot V \tag{2}$$

Here, S represents the softmax function.

– **Decoder:** During the decoding phase, every decoder layer receives the output of the encoder (the K and V matrices) and the output of the previous decoder layer. Additionally, we modified the self-attention layers into what we defined as "Masked" self-attention layers. The masked MH self-attention layer ensures the use of only the self-attention scores. We do it by adding a factor M to the word embeddings in Eq. 3. We set M to -inf for masked positions and 0 otherwise.

$$Z = S\left(\frac{Q \cdot K^T + M}{\sqrt{d_k}}\right) \cdot V \tag{3}$$

– **Preprocessing:** For performing the preprocessing, we should note that the two proposed models, BERT and DistilBert, based on deep neural network architectures, include similar steps for removing special characters, lemmatization, and stop word removal. In addition, tokenized documents are truncated or padded with a given number of tokens to ensure that the model receives uniformly sized input samples (i.e., with the same number of tokens).

As mentioned above, we will develop the problem using pre-trained BERT and DistilBERT for automatic news categorization and test different optimizer methods. Table 1 shows the parameters used for each architecture.

Table 1. Hyperparameters used in both models

BERT		DistilBert	
vocab size	128000	vocab size	128000
hidden size	768	hidden size	768
num hidden layers	12	num hidden layers	6
num attention heads	12	num attention heads	12
intermediate size	3072	intermediate size	3072
hidden dropout prob	0.1	hidden dropout prob	0.1
attention probs	0.1	attention probs	0.1
dropout prob	0.1	dropout prob	0.1
max position embeddings	512	Seq classif dropout	0.2
type vocab size	2	type vocab size	2
initializer range	0.02	initializer range	0.02
interaction	32	interaction	20
Optimizer	Adam Optimizer	Optimizer	Adamax Optimizer

4 Experimental Results

We performed testing on 55% of the test data as suggested by the competition parameters set by National Yang Ming Chiao Tung University (NYCU).

4.1 Machine Learning Results

Following the feature extraction process outlined in the preceding section, we evaluated each classifier with various feature extraction methods, and the results are in Table 2. As anticipated, Word2Vec emerged as one of the feature extraction models contributing significantly to the model's overall performance. The best

results among the three proposed machine learning models, according to the predefined metrics, were achieved using Word2Vec.

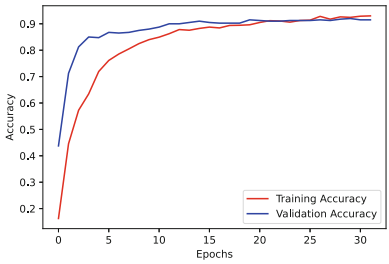
The top-performing machine learning model was the Support Vector Machine (SVM), employed based on the premise that it worked well with Word2Vec. During development, we tested three kernel types: 'linear,' 'poly,' and 'rbf,' to find the best approximation to align with the probability estimates derived from Word2Vec. The Radial Basis Function (RBF) Kernel was the most suitable task. Nevertheless, it is noteworthy to mention that prior expectations suggested that the linear Kernel would yield superior classifier performance.

Table 2. Performance of the Classifiers

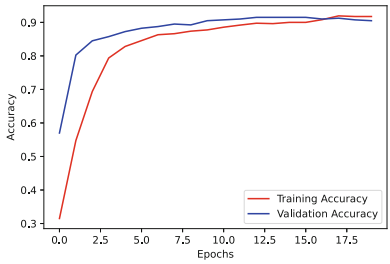
Algorithm	Feature Extraction	Accuracy	Precision	Recall	F1
Support Vector Machine	Word2Vect	0,88	0,88	0,88	0,88
LogisticRegression	Word2Vect	0,86	0,86	0,86	0,86
RandomForest	Word2Vect	0,86	0,86	0,87	0,86
Support Vector Machine	TF-IDF	0,83	0,83	0,83	0,82
LogisticRegression	TF-IDF	0,82	0,82	0,82	0,82
LogisticRegression	Bag-of-Words	0,81	0,81	0,81	0,81
Support Vector Machine	Bag-of-Words	0,8	0,8	0,81	0,8
RandomForest	TF-IDF	0,75	0,76	0,75	0,75
RandomForest	Bag-of-Words	0,74	0,76	0,74	0,74

4.2 Transformers Results

The previously described construction of the BERT and DistilBERT models resulted in two models with accuracies of 0.92 and 0.917, respectively. Since we conducted both the analysis and training, accuracy and loss are the primary metrics in Figs. 4 and 5. The accuracy trends during training for each iteration are observable, with each model utilizing a different number of iterations. Specifically, the BERT model employed 32 iterations, while DistilBERT used 20. To assess the efficiency of each model in achieving their maximum accuracy levels, we implemented EarlyStopping with a patience of 5 and obtained this information, as presented in Table 3. It is worth noting that while the BERT model achieves higher accuracy than DistilBERT, considering the number of iterations it takes to reach maximum accuracy is crucial. Based on the duty cycle percentage required to attain its maximum accuracy, DistilBERT emerges as the more efficient model, achieving similar accuracy to BERT in a shorter duty cycle. However, it is noteworthy to comment on the significance of this observation despite the highest accuracy.

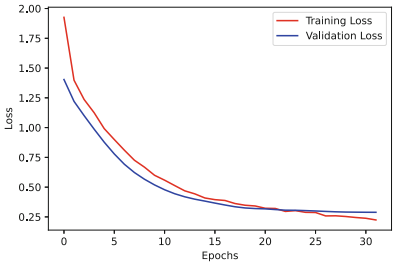


(a) Accuracy BERT model with 32 spochs.

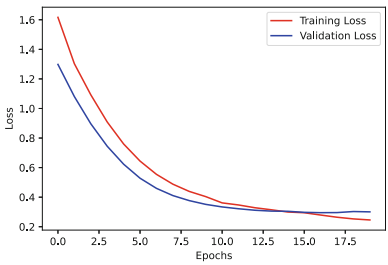


(b) Accuracy DistilBERT model with 20 spochs.

Fig. 4. Training and validation accuracy.



(a) Loss BERT model with 32 spochs.



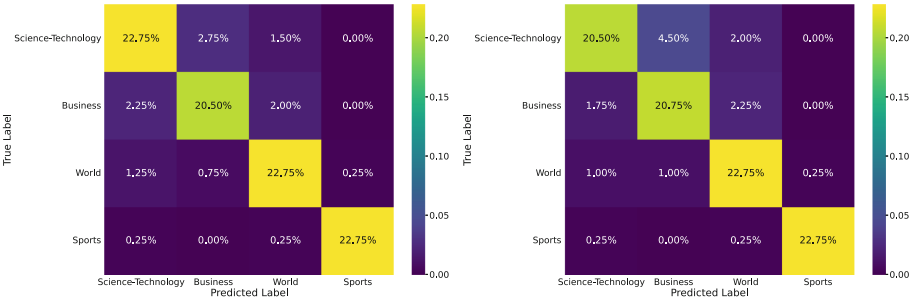
(b) Loss DistilBERT model with 20 spochs.

Fig. 5. Training and validation loss.

In Fig. 6a, it is evident that DistilBERT does not achieve equal or higher precision than BERT. However, it excels in predicting the science/technology and business categories with a minimum percentage error of 2.25% when considering the dataset. Furthermore, DistilBERT is highly effective in predicting the sports category, which comprises a significantly larger volume of data than the other classes, namely science/technology and business.

Table 3. Efficiency comparison measured over the duty cycle.

Model	interactions	Loss	Accuracy	Duty cycle
BERT	30/32	0.2889	0.92	93.17 %
DistilBERT	15/20	0.3045	0.917	75 %



(a) BERT model (b) DistilBERT model

Fig. 6. Transformer model confusion matrices for news classification.

5 Conclusions

The study's primary objective was to find an optimal model for classifying multi-class texts into four specific categories: Science/Technology, World, Sports, and Business, addressing a challenge posed by National Yang Ming Chiao Tung University (NYCU). The approach involved comparing results from three machine-learning models and two Transformer models.

Machine Learning models explored various feature extraction methods to enhance news classification accuracy. The top-performing model was Support Vector Machine in conjunction with Word2Vec, achieving an accuracy of 88%, surpassing other models. Word2Vec consistently proved to be the best feature extractor, enhancing accuracy by 1% to 2% across models, with sports news classification showing the highest precision and AUC.

Two Transformer models, BERT and DistilBERT, were also evaluated. DistilBERT emerged as the more efficient model, offering faster and more accurate news classification despite having slightly lower accuracy than BERT. Ultimately, DistilBERT was the preferred classifier with an accuracy of 91.7%, outperforming other models.

Comparing the results with 58 other participants, the Machine Learning model ranked seventh, while the Transformer model claimed the top position with a marginal accuracy difference of 0.01%.

For future work, there are plans to optimize the model further. Additionally, a multilingual model capable of classifying news in English, Mandarin, Spanish, and Hindi, the world's most spoken languages according to the BBC, will be developed.

References

1. lab 912, M.: Deeplearning hw2 transformer (2022). https://kaggle.com/competitions/deeplearning-hw2-transformer

2. Ahmed, J., Ahmed, M.: Online news classification using machine learning techniques. IIUM Eng. J. **22**, 210–225 (2021). https://doi.org/10.31436/iiumej.v22i2.1662, https://journals.iium.edu.my/ejournal/index.php/iiumej/article/view/1662
3. Ahmed, J., Ahmed, M.: Online news classification using machine learning techniques. IIUM Eng. J. **22**, 210–225 (2021). https://doi.org/10.31436/iiumej.v22i2.1662, https://journals.iium.edu.my/ejournal/index.php/iiumej/article/view/1662
4. Patro, A., Mahima Patel, R.S., Save, D.J.: Real time news classification using machine learning. Int. J. Adv. Sci. Technol. **29**(9s), 620–630 (2020)
5. Barua, A., Sharif, O., Hoque, M.M.: Multi-class sports news categorization using machine learning techniques: resource creation and evaluation. Procedia Compute. Sci. **193**, 112–121 (2021). https://doi.org/10.1016/j.procs.2021.11.002, https://www.sciencedirect.com/science/article/pii/S1877050921021268. 10th International Young Scientists Conference in Computational Science, YSC2021, 28 June–2 July 2021
6. Blackledge, C., Atapour-Abarghouei, A.: Transforming fake news: robust generalisable news classification using transformers (2021). https://doi.org/10.48550/ARXIV.2109.09796, http://arxiv.org/2109.09796
7. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation (2014). https://doi.org/10.48550/ARXIV.1406.1078, http://arxiv.org/1406.1078
8. Deb, N., Jha, V., Panjiyar, A., Gupta, R.: A comparative analysis of news categorization using machine learning approaches. Int. J. Sci. Technol. Res. **9**, 2469–2472 (2020)
9. Devi, J.S., Bai, D.M.R., Reddy, C.: Newspaper article classification using machine learning techniques. Int. J. Innov. Technol. Explor. Eng. **9**(5), 872–877 (2020). https://doi.org/10.35940/ijitee.e2753.039520, https://dx.doi.org/10.35940/ijitee.E2753.039520
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding (2018). https://doi.org/10.48550/ARXIV.1810.04805, http://arxiv.org/1810.04805
11. Elnagar, A., Al-Debsi, R., Einea, O.: Arabic text classification using deep learning models. Inf. Process. Manag. **57**(1), 102121 (2020). https://doi.org/10.1016/j.ipm.2019.102121, https://www.sciencedirect.com/science/article/pii/S0306457319303413
12. Gillioz, A., Casas, J., Mugellini, E., Khaled, O.A.: Overview of the transformer-based models for NLP tasks. In: 2020 15th Conference on Computer Science and Information Systems (FedCSIS), pp. 179–183 (2020). https://doi.org/10.15439/2020F20
13. Greff, K., Srivastava, R.K., Koutnik, J., Steunebrink, B.R., Schmidhuber, J.: LSTM: a search space odyssey. IEEE Trans. Neural Netw. Learn. Syst. **28**(10), 2222–2232 (2017). https://doi.org/10.1109/tnnls.2016.2582924
14. Kosheleva, O., Kreinovich, V., Shahbazova, S.: Type-2 fuzzy analysis explains ubiquity of triangular and trapezoid membership functions. In: Shahbazova, S.N., Kacprzyk, J., Balas, V.E., Kreinovich, V. (eds.) Recent Developments and the New Direction in Soft-Computing Foundations and Applications. SFSC, vol. 393, pp. 63–75. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-47124-8_6
15. Lilleberg, J., Zhu, Y., Zhang, Y.: Support vector machines and word2vec for text classification with semantic features. In: 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), pp. 136–140 (2015). https://doi.org/10.1109/ICCI-CC.2015.7259377

16. Luo, X.: Efficient English text classification using selected machine learning techniques. Alex. Eng. J. **60**(3), 3401–3409 (2021). https://doi.org/10.1016/j.aej.2021.02.009, https://www.sciencedirect.com/science/article/pii/S1110016821000806
17. Munikar, M., Shakya, S., Shrestha, A.: Fine-grained sentiment classification using BERT. In: 2019 Artificial Intelligence for Transforming Business and Society (AITB), vol. 1, pp. 1–5 (2019). https://doi.org/10.1109/AITB48515.2019.8947435
18. Pennington, J., Socher, R., Manning, C.: GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. Association for Computational Linguistics, Doha (2014). https://doi.org/10.3115/v1/D14-1162, https://www.aclanthology.org/D14-1162
19. Qadi, L.A., Rifai, H.E., Obaid, S., Elnagar, A.: Arabic text classification of news articles using classical supervised classifiers. In: 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS), pp. 1–6 (2019). https://doi.org/10.1109/ICTCS.2019.8923073
20. Rustamov, S.: A hybrid system for subjectivity analysis. Adv. Fuzzy Syst. **2018**, 1–9 (2018). https://doi.org/10.1155/2018/2371621
21. Rustamov, S., Mustafayev, E., Clements, M.: Context analysis of customer requests using a hybrid adaptive neuro fuzzy inference system and hidden Markov models in the natural language call routing problem. Open Eng. **8**, 61–68 (2018). https://doi.org/10.1515/eng-2018-0008
22. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter (2019). https://doi.org/10.48550/ARXIV.1910.01108, http://arxiv.org/1910.01108
23. Vaswani, A., et al.: Attention is all you need (2017). https://doi.org/10.48550/ARXIV.1706.03762, http://arxiv.org/1706.03762
24. Yang, Y., Chen, X., Tan, R., Xiao, Y.: IoT Technologies and Applications, pp. 1–60. Wiley (2021). https://doi.org/10.1002/9781119593584.ch1
25. Yıldırım, S., Jothimani, D., Kavaklıoğlu, C., Başar, A.: Classification of "hot news" for financial forecast using NLP techniques. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 4719–4722 (2018). https://doi.org/10.1109/BigData.2018.8621903