

Analysing the Sentiments in Online Reviews with Special Focus on Automobile Market

Ayman Yafoz¹, Fariyal Syed², Malek Mouhoub²^a and Lisa Fan²

¹Department of Information Systems, King Abdulaziz University, Abdullah Sulayman Street, Jeddah, Saudi Arabia

²Department of Computer Science, University of Regina, 3737 Wascana Parkway, Regina, Canada

Keywords: Canadian Automobiles Market, Sentiment Analysis, Machine Learning, Deep Learning.

Abstract: Analysing the sentiments in online reviews assists in understanding customers' satisfaction with a provided service or product, which gives the industry an opportunity to enhance the quality of their commodity, increase sales volume, develop marketing strategies, improve response to customers, promote customer satisfaction, and enhance the industry image. However, the studies focusing on applying machine learning algorithms and word embedding models, as well as deep learning techniques to classify the sentiments in reviews extracted from automobile forums, are arguably limited, and to fill this gap, this research addressed this area. Moreover, the research concentrated on categorizing positive, negative, and mixed sentiment categories in online forum reviews. The procedures for gathering and preparing the dataset are illustrated in this research. To perform the classification task, a set of models which include supervised machine learning, deep learning, and BERT word embedding is adopted in this research. The results show that the combination of the BERT word embedding model with the LSTM model produced the highest F1 score. Finally, the paper lays out recommendations to enhance the proposed system in future studies.


1 INTRODUCTION

Sentiment analysis has recently gained an increasing amount of attention from researchers due to its importance as well as the growth of social media. However, the current contributions addressing sentiment analysis on online reviews about automobiles are arguably not adequate (Wijnhoven et al., 2017). Furthermore, many phrases that appear in automobile reviews are solely related to the automobile industry and are not frequently used in other reviews. For example, the phrase "It has more ground clearance than majority of CUVs." conveys a positive sentiment, while the phrase "Massive Hyundai Engine Recall" conveys a negative sentiment. These factors motivated us to conduct this work and also to limit the scope of this research to online automobile reviews.

Moreover, this research is targeting the analysis of the sentiments into positive, negative, or mixed categories to provide a fine-grained classification of sentiments beyond the classical coarse-grained classification that is limited to only negative and

positive sentiments. The data in this research was gathered from a Canadian online forum specializing in automobile topics called Autos.ca (Autos.ca). This work addresses the lack of a customized sentiment analyser working on text exclusively about Canadian automobiles with reviews written mostly in Canadian English. The dataset has been uploaded on GitHub (named English Automobile Dataset) and can be freely accessed by future researchers, who only intend to use it for academic purposes, through the link in (English Automobiles Dataset).

The code was written in Python due to its large community, ease of use, libraries (for instance, Pandas, NLTK, and Sklearn), and the language-adequate handling of sentiment analysis tasks. On the other hand, MySQL Server was utilized to save the dataset as it is compatible with both the Python environment and Windows operating system. Moreover, an NVIDIA Tesla P100 GPU with Google Cloud was utilized to train the deep learning and word embedding models. The Keras neural networks API and TensorFlow platform were utilized to provide the deep learning libraries.

 <https://orcid.org/0000-0001-7381-1064>

The remaining sections of this paper describe the different steps of our proposed methodology and are divided as follows. Section 2 provides a literature review discussing several sentiment analysis contributions. Section 3 explains in detail the data assembly and annotation phases. Section 4 highlights the phases of the pre-processing, while Section 5 shows the phases of the feature selection. Section 6 illustrates the phases of splitting and balancing the dataset. Section 7 lists and analyses the outcomes of the machine learning models. Section 8 overviews the BERT word embedding model as well as the deep learning models, and reports on the related experimental results. Section 9 focuses on the data visualization technique. Finally, Section 10 provides a list of concluding remarks and future enhancements to improve the quality of this research work.

2 LITERATURE REVIEW

(Yafoz et al., 2021) analysed the sentiments in datasets containing Arabic online reviews about Arabic real estate and automobiles. They employed the BERT word embedding model with a set of deep learning algorithms: BiLSTM (Bidirectional Long Short-Term Memory), LSTM (Long Short-Term Memory), GRU (Gated Recurrent Unit), CNN (Convolutional Neural Networks), and CNN-GRU. The automobile dataset had almost 6,585 opinions, while the real estate dataset contained almost 6,434 opinions. The records in both datasets were split into three sentiment types (negative, positive, and mixed). For the dataset concerning automobiles, the highest F1 score was 98.71% using the BERT model with the LSTM. On the other hand, the highest obtained F1 score for the real estate dataset was 98.67% using the BERT model with the CNN.

(Malik et al., 2018) performed a sentiment analysis on a dataset that had 2,000 reviews divided evenly between positive and negative reviews. These reviews were written in Roman Urdu about automobiles. For the classification process, eight classifiers were utilized: bagging, Multinomial Naïve Bayes, AdaBoost, Random Forest, SVM, Deep Neural Network, Decision Tree, and K Nearest Neighbor. The highest accuracy result achieved by the Multinomial Naïve Bayes classifier was 89.75%.

(Alsawalqah et al., 2015) classified the sentiments in automobile tweets concerning three automobile manufacturers (BMW, Audi, and

Mercedes). The dataset contained 3000 tweets divided equally among the three automobile manufacturers (1000 tweets for each automobile manufacturer). The researchers used the Naïve Bayes algorithm to classify the sentiments. The results reflected that Audi had the lowest negative polarity (only 16%) and the highest positive polarity (around 83%) compared with Mercedes and BMW. This reflected that the reviewers were more satisfied with Audi than with Mercedes and BMW.

Finally, (Yafoz et al., 2020) classified three sentiment categories (mixed, positive and negative) in Arabic online reviews concerning automobiles and real estate. The dataset of real estate included 6,434 opinions, while the automobile dataset included 6,585 opinions. The researchers applied twenty-two machine learning, four word embedding algorithms (Fasttext, Glove, CBOW, and Skip-gram), and four deep learning models (LSTM, GRU, CNN, and BiLSTM) to classify the reviews. For the Arabic automobile dataset, the highest F1 score was 84.90% by the CBOW with the GRU. On the other hand, for the dataset concerning real estate, the highest F1 score was 71.33% with the combination of the Skip-gram with the GRU and CNN.

3 DATA ASSEMBLY AND ANNOTATION

The purpose of assembling and labelling the records of the dataset is to create a dataset that is suitable for supervised classification.

3.1 Data Assembly

The Octoparse web crawler (Octoparse Web Scraper) was used to extract and organize the data due to its simplicity, quickness, efficiency, and ability to extract and organize the data. The dataset source is Autos.ca, which is a Canadian automobile online forum. The dataset size is of 4014 unique records where all duplicated records were removed using the option of remove duplicates offered by Microsoft Excel. The dataset was divided into 2078 positive, 1467 negative, and 469 mixed-opinion records, and it focused on almost 56 domains related to automobiles. Table 1 shows an example of analysing the sentiments in reviews from the dataset.

Table 1: An Example of a Sentiment Analysis Performed on the Dataset.

Review	Sentiment
I do not like the Civics. Too slow for my taste.	Negative
I love the look of the current Mustang. It is certainly reminiscent of the 60s and 70s.	Positive
Bricklin SV1 had a lot of forward-thinking safety and performance features. Some good, some bad.	Mixed

3.2 Data Annotation

In this research, the manual labelling approach was adopted as it yielded a more elevated degree of accurate labelling compared to other approaches (Heo et al., 2021). Three annotators performed the labelling, and the inter-rater agreement was evaluated through the Fleiss Kappa index to assess both the reliability and consistency of the annotators. Fleiss Kappa is the most widely implemented Kappa in labelling emotions (Podlesek et al., 2009). The minimum degree of inter-rater agreement that is acceptable by many researchers is 80% (McHugh, 2012). In this research, the calculated Fleiss Kappa degree for the inter-rater agreement was 96.8%, which denotes almost perfect agreement.

4 PRE-PROCESSING

Pre-processing is the main operation in sentiment analysis tasks (Awajan et al., 2018). If pre-processing operations are not performed, it could lead the analyser to override significant terms. However, overuse of pre-processing approaches could lead to a loss of significant data (Mansour et al., 2017). In this research, the pre-processing operations were divided into two: normalization, and removal of stop words.

4.1 Normalization

Text extracted from social media is usually not ready for language processing tasks as it is written using highly informal language. This informal text needs to be normalized to an acceptable standard formal style (Erianda et al., 2017). Hence, the text in this research was automatically normalized to clean the data, remove the usernames of the reviewers to ensure privacy and anonymity, remove punctuation, exclude non-printable ASCII characters, remove non-English letters, and convert words from uppercase to lowercase to reduce the uncertainty that could be

entailed in the classification process. The following example illustrates the normalization process:

The sentence before normalization: “You bought a \$1200 twenty six year old BMW. Expect plenty of repairs on an ongoing basis”.

The sentence after normalization: “you bought a twenty six year old bmw expect plenty of repairs on an ongoing basis”.

4.2 Removing Stop Words

In many cases, stop words are useless for processing. Hence, they are discarded to save both size and time (Awajan et al., 2018). Therefore, a file composed of around 97 English stop words gathered from (Default English Stop Words List) was created by us. The following example illustrates the operation:

The original sentence: “Personally, I think it is absolutely hideous. My eyes definitely do not see any elegance with that car”.

The sentence after removing stop words: “Personally think absolutely hideous eyes definitely not see elegance car”.

5 FEATURE SELECTION

In this research, the feature selection operations were conducted to decrease the dimensionality of the dataset by reducing the initial features and retaining only important features for classification (Alonso-Betanzos et al., 2015). Four widely applied feature selection techniques were utilized, which are: N-Gram Feature, Lemmatization, POS Tagger (Part-Of-Speech Tagger), and TFIDF (Term Frequency-Inverse Document Frequency).

TextBlob Lemmatizer was utilized to lemmatize the data. It was utilized due to its high accuracy in producing lemmas, and because it is compatible with TextBlob POS tagger which was also utilized in this research to produce POS tags. For illustration, the following instance demonstrates a lemmatization operation performed by TextBlob Lemmatizer.

The original sentence: “My wife loves the looking of Audi A8. She did put her feet on the gas pedal to enjoy the engine’s sound.”

The sentence post-lemmatization: “My wife love the look of Audi A8 She do put her foot on the gas pedal to enjoy the engine’s sound.”

TextBlob was also adopted to carry out the POS tagging operation as it is fast, easy to use, accessible, has the highest code quality “L5” (granted by Lumnify), and holds MIT license (Varma et al., 2018). Moreover, when we compared three widely

used English POS taggers (spaCy, WordNet, and TextBlob), the accuracy of identifying and tagging prepositions, pronouns, and the possessive ending was the highest with Textblob POS tagger. Table 2 illustrates how a sentence from our dataset is tagged by TextBlob POS Tagger. The sentence is “I like the price on that Van.”

Table 2: Example of a Sentence Tagged by TextBlob POS Tagger.

Word	POS Tag	Meaning (Penn Treebank II Tag Set)
I	PRP	pronoun, personal
Like	VBP	verb, non-3 rd person singular present
The	DT	determiner
Price	NN	noun, singular, or mass
On	IN	conjunction, subordinating, or preposition
That	DT	determiner
Van	NNP	noun, proper singular

6 SPLITTING AND BALANCING THE DATASET

The dataset in this research was prepared in three phases: splitting the dataset between training (70%) and testing (30%), 10 K-fold cross-validation, and oversampling the minority classes (the SMOTE-NC “Synthetic Minority Oversampling Technique Nominal and Continuous” was used to conduct synthetic oversampling operation over the classes with minority occurrence in the training datasets).

7 THE SUPERVISED MACHINE LEARNING APPROACH

Five machine learning classifiers were used in this research, which are Linear SVC, Bernoulli Naïve Bayes, the Multi-layer Perceptron (MLP) Classifier, CART Decision Tree, and Multinomial Naïve Bayes. Additionally, two classifiers were also utilized, which are the Ensemble Vote classifiers (hard and soft) as shown in Table 3 (Alsafari et al., 2021). The hyperparameter tuning was conducted to choose the optimum hyperparameters for the classifiers rendering the best scores when tested on the training dataset. Moreover, a study conducted by (Bergstra et al., 2012) showed that random search theoretically

and empirically resulted in better outcomes when optimizing hyperparameters as opposed to grid search. Hence, in this research, the random search approach was applied to tune the classifiers’ hyperparameters.

Table 3: The F1 Scores of the Machine Learning Classifiers.

Model	Best Parameters and Score	F1-Score
Linear SVC	best_params: {'max_iter': 1500, 'C': 1}. Best score: 0.889066	76%
Bernoulli Naïve Bayes	best_params: {'alpha': 1.0}. best score: 0.872703	77%
MLP Classifier	best_params: {'hidden_layer_sizes': 100, 'alpha': 0.1}. best score: 0.908035	77%
CART Decision Tree	best_params: {'max_depth': 14, 'criterion': 'entropy'}. best score: 0.785935	68%
Multinomial Naïve Bayes	best_params: {'alpha': 1.0}. best score: 0.826234	68%
The Ensemble Soft Vote	NA	80%
The Ensemble Hard Vote	NA	78%

Based on the results shown in Table 3, the Ensemble Soft Vote classifier surpassed the other models by achieving 80% in F1 scores.

8 THE BERT AND DEEP LEARNING MODELS APPROACH

BERT is an advanced and modern word embedding model developed in 2018 by Google. It was developed with the purpose of pre-training deep bidirectional segments from unlabelled data through combining right and left context in every layer. Consequently, it is possible to fine-tune a pre-trained BERT model using a single extra output layer to generate advanced models for different NLP projects without making radical architecture adjustments for each project (Devlin et al., 2019). Therefore, in this research, BERT was implemented to classify the sentiments in the dataset (Alsafari et al., 2020). Moreover, deep learning yields the most optimal solutions to natural language processing tasks (Prieta et al., 2020). Therefore, in this paper, the deep learning approach was implemented in the form of the

models: CNN, the LSTM, and GRU as shown in Table 4.

The epoch value was selected to be 50 because it showed the highest results among the randomly chosen values. This value was also selected by (Lehečka et al., 2020) when they classified large-scale multi-label Wikipedia datasets. Moreover, the value of the batch size was set to be 32 because it is an adequate default value according to (Garcia-Silva et al., 2020). Furthermore, the value of the learning rate was selected to be 5e-5, which matches the value selected by (Sun et al., 2019). The stride value was selected to be 1, which was the same value selected by (Srivastava et al., 2020), and it is the most popular value for the stride (Togashi et al., 2018). Moreover, ReLU (Rectified Linear Unit) was adopted as an activation function. ReLU was also picked by (Nie et al. 2020).

The results shown in Table 4 reflect that the combination of the BERT word embedding model with the LSTM model surpassed the combination of the BERT word embedding model with the other deep learning models by scoring 99.48% in F1-scores. In general, the combination of the BERT word embedding model with the deep learning models used in this research generated quite higher F1 scores than those achieved by the machine learning classifiers employed in this research.

Table 4: The F1-Scores Resulted From Combining BERT Word Embedding Model with a Set of Deep Learning Models.

Deep Learning Model	BERT F1-Score
CNN	99.19%
LSTM	99.48%
GRU	94.73%
BERT	98.22%

9 DATA VISUALIZATION

There is a set of sentiment words and clauses that the classifiers depend on to determine the polarity of the classification. For instance, when the Bernoulli Naïve Bayes classifier analysed the sentiments in the dataset, the following sets of negative and positive words and clauses assisted in determining the sentiment polarity of the text. As stated above, some of these sentiment words and clauses (such as recall, of power, and head gasket, among others) are rarely used outside of the automobile domain. This justifies limiting the scope of this research to automobile data as general sentiment analyzers will arguably not be

able to classify these words and clauses. Some of these words and clauses are illustrated in Figure 1 and also shown below:

Positive words and clauses:

['more fun', 'world', 'sweet', 'much good', 'it very', 'be good', 'be great', 'be much', 'white', 'genesis', 'beautiful', 'best', 'best car', 'excellent', 'awesome', 'nice', 'overall', 'look great', 'comfy', 'one of most', 'one of best', 'very nice', 'car but', 'safe', 'very comfortable', 'of power', 'fantastic', 'fine', 'of best', 'fun drive'].

Negative words and clauses:

['car but', 'terrible', 'not like', 'crap', 'break', 'unreliable', 'crappy', 'fall', 'noise', 'failure', 'uncomfortable', 'awful', 'ugly', 'certain', 'recall', 'po', 'pricey', 'fail', 'but be', 'and not', 'more expensive', 'but not', 'gasket', 'horrible', 'head gasket', 'worst', 'hate', 'hat', 'weak', 'poor'].

10 CONCLUSION AND FUTURE WORK

Despite the improvements suggested for future work outlined below, it is evident that the methodology used in this research successfully filled the gaps that were left unaddressed by other contributions regarding the analysis of sentiments that exist in reviews in the automobile domain. In terms of the results, the combination of the BERT word embedding model with the LSTM model had the highest F1-score, which reflects an opportunity for researchers to adopt such a combination to analyse the sentiments in English automobile data in particular, and non-English automobile data in general. The methodology adopted in this research has also shown superior F1 scores when compared with the scores achieved by other works that were reviewed in this paper.

In future work, adopting advanced models such as reinforcement learning, ERNIE, and Elmo could enhance the results and widen the scope of this area of research. The results could also be improved by enlarging the dataset, treating negations, and developing specific word embedding models that are more related to the automobile industry in terms of the embedded vocabulary. The scope of the work could also be broadened by covering the semi-supervised approaches (Alsafari et al., 2021). Finally, performing an aspect-based sentiment analysis would result in more precise sentiment analysis for the opinion target.

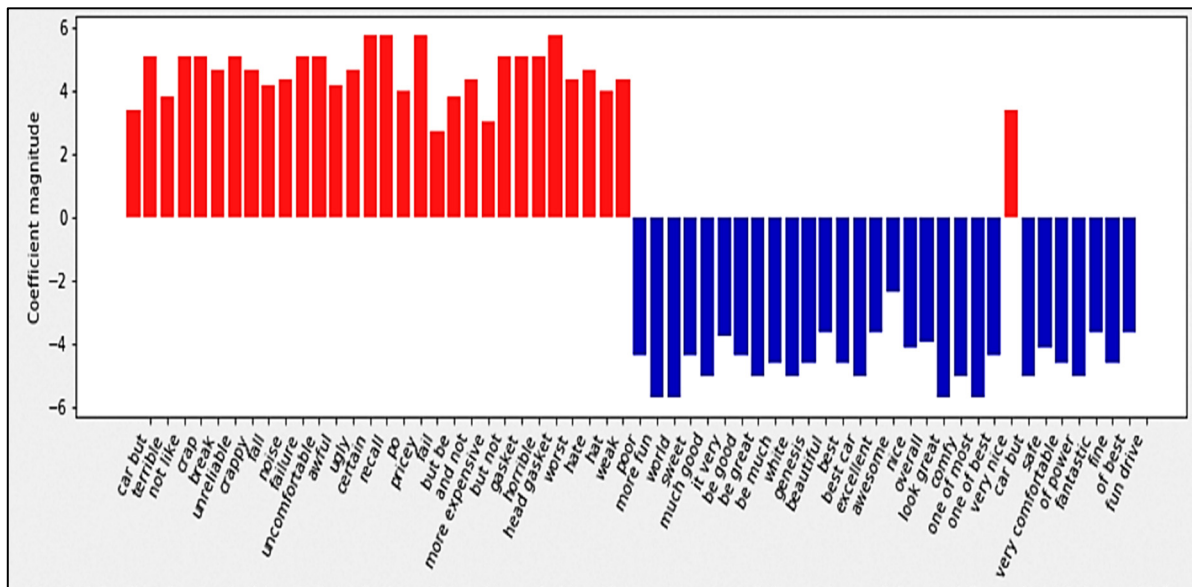


Figure 1: The Negative and Positive Words and Clauses that Assisted the Classifier in Determining the Sentiment Polarity of the Text.

REFERENCES

- Wijnhoven, F., Plant, O. (2017). Sentiment Analysis and Google Trends Data for Predicting Car Sales. *The Thirty Eighth International Conference on Information Systems*.
- Autos.ca. [cited 05-01-2021]; Available from: https://www.autos.ca/forum/index.php?board=6_7.0
- English Automobiles Dataset. [cited 24-9-2021]; Available from: <https://github.com/aymanya/English-Automobiles-Dataset>
- Yafoz, A., Mouhoub, M. (2021). Sentiment Analysis in Arabic Social Media Using Deep Learning Models. *In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2021)*.
- Malik, K. Khan, M. (2018). Sentiment Classification of Customer's Reviews about Automobiles in Roman Urdu. *In Proceedings of the 2018 Future of Information and Communication Conference (FICC)*.
- Alsawalqah, H., Aljarah, I., Yaghi, R., Shukri, S. (2015). Twitter Sentiment Analysis: A Case Study in the Automotive Industry. *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AECT)*.
- Yafoz, A., Mouhoub, M. (2020). Analyzing Machine Learning Algorithms for Sentiments in Arabic Text. *In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2020)*.
- Octoparse Web Scraper. [cited 24-02-2021]; Available from <https://www.octoparse.com/>
- Heo, G., Whang, S., Roh, Y. (2021). A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. *IEEE Transactions on Knowledge and Data Engineering Journal*. 33(4).
- Podlesek, A., Boštjan, V., Mihelič, F., Komidar, L., Gajšek, R., Štruc, V. (2009). Analysis and Assessment of AVID: Multi-Modal Emotional Database. *In International Conference on Text, Speech and Dialogue (TSD 2009)*.
- McHugh, L. (2012). Interrater Reliability: The Kappa Statistic. *The Croatian Society of Medical Biochemistry and Laboratory Medicine (Biochemia Medica) Journal*. 22(3).
- Awajan, I., Mohamad, M. (2018). A Review on Sentiment Analysis in Arabic Using Document Level. *International Journal of Engineering and Technology*. 7((3.13) (2018)).
- Mansour, H., El-Masri, M., Altrabsheh, N. (2017). Successes and Challenges of Arabic Sentiment Analysis Research: A Literature Review. *Social Network Analysis and Mining Journal*. 7(1).
- Erianda, A., Rahmayuni, I. (2017). Improvement of Email and Twitter Classification Accuracy Based on Preprocessing Bayes Naïve Classifier Optimization in Integrated Digital Assistant. *The International Journal on Informatics Visualization*. 1(2).
- Default English Stop Words List. [cited 03-10-2021]; Available from: <https://www.ranks.nl/stopwords>
- Alonso-Betanzos, A., Sánchez-Marño, N., Bolón-Canedo, V. (2015). *Feature Selection for High Dimensional Data (Artificial Intelligence: Foundations, Theory, and Algorithms)*. Springer. 1st Edition.
- Varma, A., Saha, A., Kunal, S., Tiwari, V. (2018). Textual Dissection of Live Twitter Reviews Using Naïve Bayes. *Procedia Computer Science Journal*. 132(2018).

- Penn Treebank II Tag Set. [cited 07-04-2021]; Available from: <http://relearn.be/2015/training-common-sense/sources/software/pattern-2.6-critical-fork/docs/html/mbsp-tags.html>
- Alsafari, S., Sadaoui, S. (2021). Ensemble-based Semi-Supervised Learning for Hate Speech Detection. *Proceedings of the Thirty-Fourth International Florida Artificial Intelligence Research Society Conference*.
- Bergstra, J., Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*. 13(2012).
- Devlin, J., Lee, K., Toutanova, K., Chang, M. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*.
- Alsafari, S., Sadaoui, S., Mouhoub, M. (2020). Hate and Offensive Speech Detection on Arabic Social Media. *Online Social Networks and Media Journal*. 19.
- Prieta, F., Moreno-García, M., Dang, N. (2020). Sentiment Analysis Based on Deep Learning: A Comparative Study. *Electronics Journal*. 9(3,483).
- Lehečka, J., Švec, J., Šmídl, L., Ircing, P. (2020). Adjusting BERT's Pooling Layer for Large-Scale Multi-Label Text Classification. *International Conference on Text, Speech, and Dialogue (TSD 2020)*.
- Garcia-Silva, A., Gomez-Perez, J., Denaux, R. (2020). *A Practical Guide to Hybrid Natural Language Processing: Combining Neural Models and Knowledge Graphs for NLP*. Springer.
- Sun, C., Qiu, X., Xu, Y., Huang, X. (2019). How to Fine-Tune BERT for Text Classification. *China National Conference on Chinese Computational Linguistics (CCL 2019)*.
- Srivastava, H., Srivastava, S., Kumari, S., Varshney, V. (2020). A Novel Hierarchical BERT Architecture for Sarcasm Detection. *Proceedings of the Second Workshop on Figurative Language Processing*.
- Togashi, K., Nishio, M., Do, R., Yamashita, R. (2018). Convolutional Neural Networks: An Overview and Application in Radiology. *Insights into Imaging Journal*. 9(4).
- Nie, J., Du, P., Lu, Z. (2020). VGCN-BERT: Augmenting BERT with Graph Embedding for Text Classification. *European Conference on Information Retrieval (ECIR 2020)*.
- Alsafari, S., Sadaoui, S., Mouhoub, M. (2021). Semi-Supervised Self-Training of Hate and Offensive Speech from Social Media. *Applied Artificial Intelligence Journal*.