Comparison of Word Embeddings for Sentiment Classification with Preconceived Subjectivity

Xi Jie Lee[1], Timothy Tzen Vun Yap[1(✉)], Hu Ng[1], and Vik Tor Goh[2]

[1] Faculty of Computing and Informatics, Multimedia University, 63100 Cyberjaya, Malaysia
timothy@mmu.edu.my

[2] Faculty of Engineering, Multimedia University, 63100 Cyberjaya, Malaysia

Abstract. This research looks into objectivity and subjectivity's effects on sentiment analysis through word embeddings, namely Word2Vec, Term Frequency-Inverse Document Frequency (TF-IDF), and Bidirectional Encoder Representations from Transformers (BERT). Objectivity corpora are defined as data with a neutral point of view and no biases. In contrast, subjectivity corpora are defined as data from a non-neutral point of view and may contain biases. The goals are to compare the efficacy of numerous embedding methods on sentiment analysis classification after subjectivity analysis. In terms of embedding methods, results from our work show that BERT embedding gives the best outcome for subjectivity classification with an accuracy score of 99.77%. For sentiment classification, TF-IDF provides the highest accuracy with 91.29%.

Keywords: Word2Vec · TF-IDF · BERT · sentiment analysis · word embedding · sentence embedding

1 Introduction

With the rapid growth of internet accessibility, e-commerce applications are increasing apparently. Therefore, most consumers will begin seeking product information, sharing purchase experiences, or chatting with other users on the e-commerce platform. Text mining studies have started to acquire popularity in the data analysis field. To illustrate this point, text mining may be used to extract the sentiment of product reviews, which can be essential in determining whether a product is excellent or poor and therefore impact sales directly.

Natural language processing is a subset of artificial intelligence, with emphasis on the automated discovery of the knowledge concerned with the interaction of computers and human languages. This research aims to classify sentiment after subjectivity analysis through comparison of machine learning algorithms.

The first phase of this research project is focused on the difference between objectivity and subjectivity corpus. The second phase is focused on classifying the sentiment from the subjectivity corpus, as objective corpora should not exhibit sentiment. In preparation of the classification, word and sentence embedding algorithms first employed, namely Word2Vec, TF-IDF, and BERT embedding.

S.-C. Haw and K. Sonai Muthu (Eds.): CITIC 2022, 10, pp. 488–502, 2022.
https://doi.org/10.2991/978-94-6463-094-7_39

2 Literature Review

2.1 Sentiment Analysis

Sentiment analysis is the management of emotions, opinions, and subjective writings. It is a subset of natural language processing, and its precursory area falls in artificial intelligence (Aung and Myo 2017). With big data, sentiment analysis provides understanding of information related to public opinion because of the prevalence of social media content such as tweets, comments, etc. Aside from that, with the rise of e-commerce applications, consumers are becoming accustomed to shopping online and leaving feedback on merchants about their purchase experiences. These contents are a valuable resource for future customer decisions and for merchants to improve products and services. Reviews of products can be classified as negative, positive, or neutral in terms of sentiment, and these can be used to improve customer satisfaction.

2.2 Sentiment Analysis Based on Sentiment Lexicon

The lexicon-based approach does not require any training data and sentiment is classified by a sentence that is inferred by the polarity of the words. In the instance of a sentence, the polarity of the individual words that make up the phrase together are expressed as the final sentiment of the sentence. As a result, a sentence's polarity is the sum of the polarities of the individual words in the sentence.

In Aung and Myo (2017), the researchers mapped each word with the predefined lists of databases in order to compute the polarity of the sentiment. Further, this approach can also be employed with the following techniques:

i. Dictionary-based methods: dictionaries such as WordNet are used to indicate the positive or negative of the words.
ii. Corpus-based methods: a large corpus of words is used and based on syntactic patterns.

Table 1 shows sample words in a sentiment word database, with scores that can be used to compute the sentiment result.

Figure 1 shows the process flow of sentiment analysis based on the lexicon approach. This approach consists of five steps, starting with data collection. Text processing reduces the inconsistencies and is followed by assignment of scores based on the polarity of the words, and the final sentiment is computed from the scores.

2.3 Sentiment Analysis Based on Machine Learning

Soumya and Pramod (2020) proposed an experiment in sentiment analysis in Malayalam Tweets. The researchers classified the tweets using different machine learning classifiers such as Naïve Bayes, Support Vector Machine, and Random Forest. In addition, the word embedding methods proposed by the researchers were Bag-of-Words (BOW) and TF-IDF. The challenge faced by this researcher is the unavailability of sentiment-tagged corpus to Malayalam.

Table 1. Sample words in sentiment word database (Aung and Myo 2017)

Example of Opinion Words		
Opinion Word	Score	Description
care	+2	verb
complex	−3	adjective
normal	0	adjective
daily	0	adjective
most	+100%	Intensifier
little	−50%	Intensifier

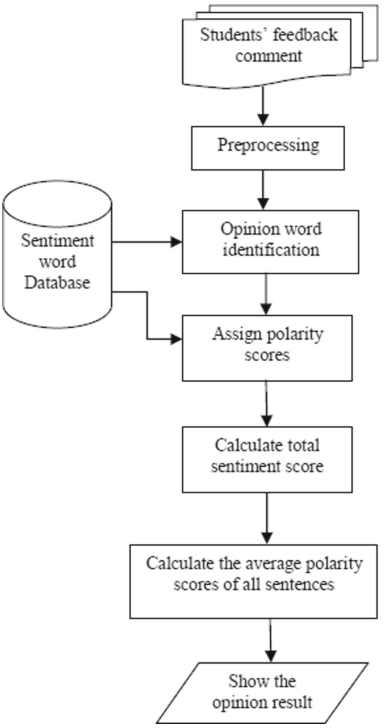

Fig. 1. Sentiment analysis based on the lexicon approach (Aung and Myo 2017)

Figure 2 shows the accuracy of the suggested approach. All the unique words in the corpus are considered while creating feature matrices using BOW and TF-IDF. Some terms are unimportant in sentiment analysis for predicting positive and negative sentiment. When comparing BOW and TF-IDF to the other two characteristics, the feature matrix with BOW and TF-IDF is larger. Because emotion-oriented words are crucial for predicting the sentiment of sentences, all three classifiers, Unigram with

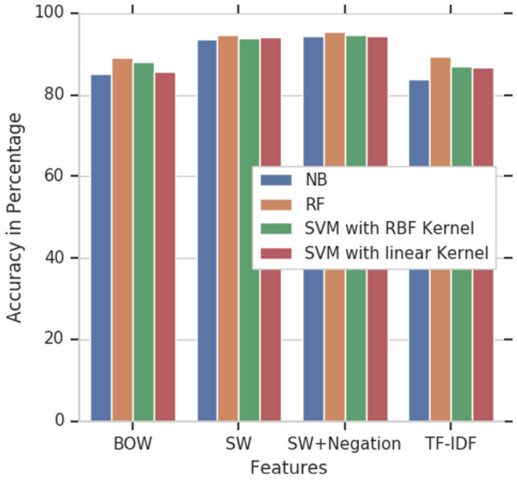

Fig. 2. Sentiment analysis results (Soumya and Pramod 2020)

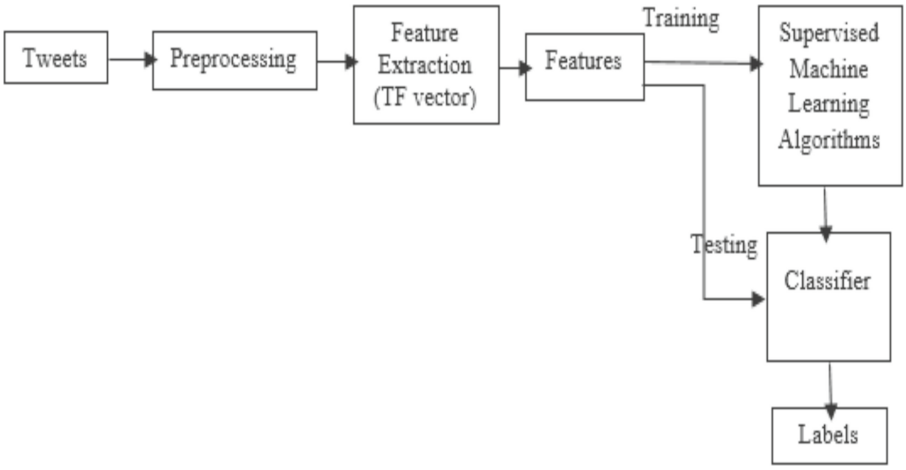

Fig. 3. Sentiment analysis based on supervised machine learning

Sentiwordnet and Unigram with Sentiwordnet adding negation words, exhibit superior accuracy.

Poornima and Priya (2020) applied term frequency-based approaches to classify the sentiment of Twitter reviews. The machine learning classifier such as Support Vector Machine, Multinomial Naïve Bayes, and Logistic Regression machine learning algorithms are used to perform sentiment analysis classification.

Figure 3 shows the process flow of sentiment analysis based on the machine learning approach. This approach consists of seven steps, starting with data collection, followed by text processing, feature extraction, features, training with supervised machine learning algorithms, testing, and last prediction for the sentiment.

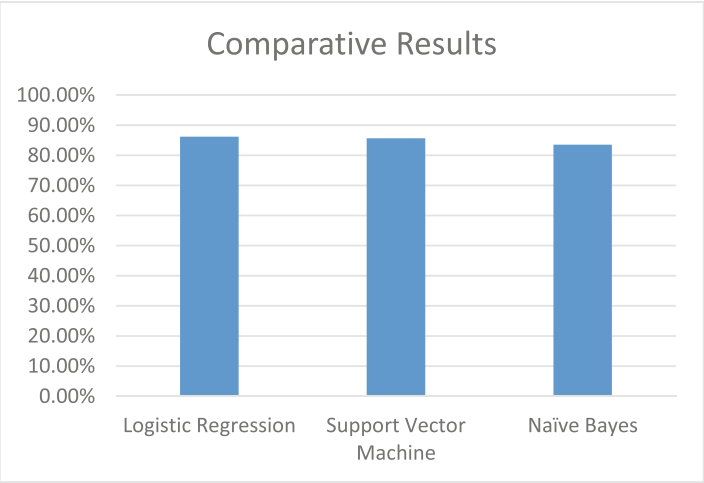

Fig. 4. Sentiment analysis classification accuracy Poornima and Priya (2020)

In this approach, feature extraction is one of the steps in developing a machine learning text classifier. Figure 4 above shows the accuracy of the proposed methods by Poornima and Priya (2020). According to Soumya and Pramod (2020) and Poornima and Priya (2020), the authors make use of feature extraction which can transform and pre-process text into numerical attributes.

2.3.1 Lexical Features

BOW is one of the lexical techniques for text classification. When applying this approach, each word becomes a feature that is assigned a weight value. The TF-IDF value of the word in the corpus is used to compute the weight of the word.

$$TF(x) = \frac{number\ of\ times\ term\ x\ appears\ in\ a\ document}{total\ number\ of\ terms\ in\ the\ document} \tag{1}$$

$$IDF(x) = \log_e \frac{total\ number\ of\ documents}{number\ of\ documents\ with\ term\ x\ in\ it} \tag{2}$$

2.3.2 Grammatical Features

Negation is one of the grammatical characteristics of sentiment analysis. The existence of a negative word can completely alter the polarity of a sentence. An example of this is "I don't really like this product". The word "like" is positive in this sentence, but the negation word "not" changes the polarity of the statement to negative. As a result, in sentiment analysis, negation handling is critical.

2.3.3 Word Embedding

Word embeddings are vector representations of underlying words that capture their context regarding other words in the phrase. This transformation will result in words of

similar meaning being grouped closer together and different words positioned further away in the hyperplane.

2.3.4 Unsupervised Machine Learning

The unsupervised machine learning approach utilizes one or more effective dictionaries to calculate the sentiment level of the sentence. These predefined dictionaries contain a list of words that have been labelled with their corresponding level of positiveness or negativeness.

2.4 Word2Vec

Word2Vec is a word embedding in the form of vector representations of underlying words that capture their context concerning other words in the phrase. According to Ahuja et al. (2019), Word2Vec works similarly to the human brain in that it employs word association to assist a computer in identifying probable word combinations. The central concept is that word embeddings will eliminate the need to build features based on stylometry to assess sentiment manually.

The texts are cleaned in a list of sentences in the Word2Vec model, and the tokenizer divides the words, letters, and symbols in each sentence. Word2Vec encompasses the continuous bag-of-words model (CBOW) and the skip-gram (SG) model. They both work well in allowing neural networks to learn words and their context. The CBOW approach predicts the next word based on the context, whereas the SG model determines the context based on the word.

To categorize texts after training the Word2Vec vectors, Naive Bayes, Logistic Regression, and Support Vector Machine techniques were utilized Munikar et al. (2019). With the Support Vector Classifier (SVC) with SG as the Word2Vec training model, the maximum accuracy rate achieved by the classifier was 72%.

2.5 TF-IDF

TF-IDF is a well-known approach for determining the significance of a word in a document (Eqs. 1 and 2). The number of times a term appears in a text divided by the total number of words in the document yields the frequency of the term, *t*. The Inverse Document Frequency (IDF) method determines a term's relevance. Some phrases, such as "is", "an", and "and", are widely used yet have little meaning.

Ahuja et al. (2019) claimed the TF-IDF word level (Term Frequency-Inverse Document Frequency) performance for sentiment analysis is 3–4% higher than using N-gram features such as Word2Vec. The sentiment analysis experiment is performed using six classification algorithms: Decision Tree, Support Vector Machine, KNN, Random Forest, Logistic Regression, and Naïve Bayes.

Figure 5 shows the proposed methodology by Ahuja et al. (2019). The first technique was data pre-processing on the dataset and extracting the features using N-grams and TF-IDF technique. After feature extraction, the researcher applied the machine learning classifiers and evaluated the results.

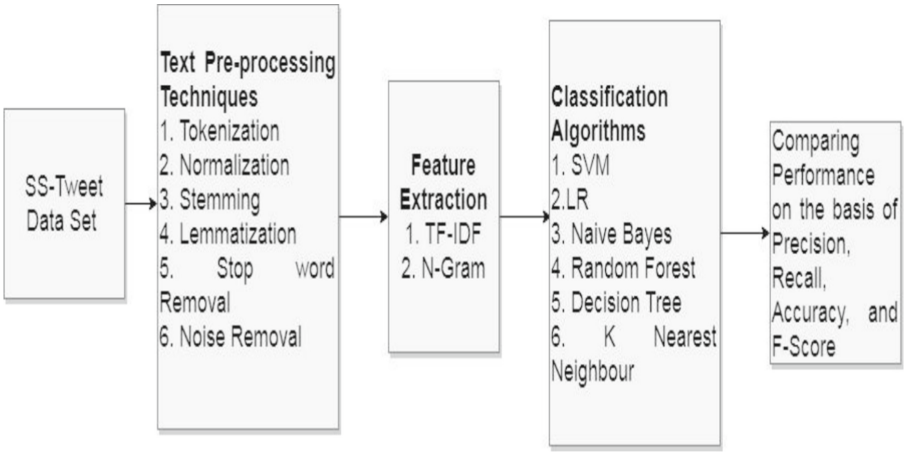

Fig. 5. The methodology for sentiment analysis (Ahuja et al. 2019)

Ahuja et al. (2019) applied six different classification algorithms to the SS-Tweet dataset considering two features (TF-IDF and N-Grams). The researchers concluded that TF-IDF features are giving better results (3–4%), shown in Table 2, when compared to N-Gram features, shown in Table 3. Besides the feature extraction, the authors also stated that Logistic Regression performed best in sentiment predictions by providing the highest output for all four comparison measures – accuracy, recall, precision, and f-score – as well as both feature extraction approaches – N-Gram and word-level TF-IDF.

2.6 Sentiment Analysis Based on Deep Learning

Machine learning algorithms are implemented to learn and act by understanding labelled input data. On the other hand, deep learning networks do not require human involvement since multilayer layers in neural networks organize input into a hierarchy of concepts that eventually learn from mistakes. The main thing affecting accuracy is always the data quality.

2.7 Bert

BERT, a state-of-the-art machine learning model is a pre-trained model that is used for sentiment analysis. According to Ain et al. (2017), BERT can extract more context features from a sequence instead of training left and right separately and the researchers mentioned that the BERT is achieved using modified language model masks known as MLM. MLM (Modified Language Masks) is to mask a random word in a sentence with a small probability.

Table 2. Result of TF-IDF Embedding (Ahuja et al. 2019)

SS-Tweet Dataset – (WORD LEVEL TF-IDF)				
ML. Algorithms	Accuracy (%)	Precision (%)	Recall (%)	F-Score (%)
KNN	46	62	33	21
Decision Tree	46	43	42	42
SVM	46	15	33	21
Logistic Regression	57	57	50	50
Naïve Bayes	53	56	44	42
Random Forest	51	47	44	44

Table 3. Result of N-Gram Embedding Ahuja et al. (2019)

SS-Tweet Dataset – (WORDLEVEL N-Gram)				
ML. Algorithms	Accuracy (%)	Precision (%)	Recall (%)	F-Score (%)
KNN	46	41	34	26
Decision Tree	48	44	42	42
SVM	46	15	33	22
Logistic Regression	49	51	39	54
Naïve Bayes	50	52	41	38
Random Forest	51	49	42	42

3 Theoretical Framework and Research Methodology

3.1 Dataset

Figure 6 shows the sentence embedding layer in BERT. In Munikar et al. (2019), the authors conducted an experiment that used the BERT model to build a sentiment classifier. BERT makes use of the modified language model mask where the entire sentence is fed into a transformer and then the model is used to predict the word. While using the BERT model, the input tokens are processed in parallel, as the whole sequence is processed at once through its attention mechanism enabled architecture.

This paper aims to compare the performance of feature vectors from objectivity and subjectivity corpora. We assume the following:

(a) Corpus is objectivity based or subjectivity based only.
(b) Corpus must be relatively large

In general, two datasets are proposed and used to perform the sentiment analysis classification. For word embedding training, the datasets shown in Table 4 are used:

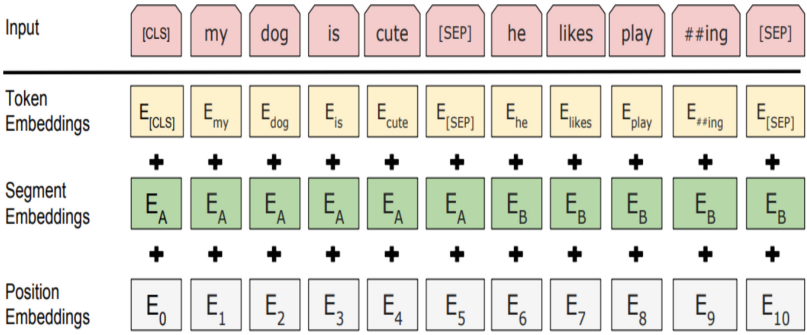

Fig. 6. BERT sentence embedding layer Munikar et al. (2019)

Table 4. Information of datasets

Wikipedia Article Dataset (Objectivity Corpus)	Shopee Product Reviews (Subjectivity Corpus)
Articles from Wikipedia are used as the objectivity dataset (King 2017)	Product reviews from the Shopee Code League 2020 are used as the subjectivity dataset (Suitnatsnoc 2020)

3.1.1 Objectivity Corpus: Embedding Data

Wikipedia is a dataset that meets the objectivity requirement. Wikipedia is a free online encyclopedia that is available in various languages. It is built on open online cooperation and has become the most comprehensive online reference on the World Wide Web (WWW). According to the Wikipedia Policy, all Wikipedia content must be written from a neutral point of view (NPOV). The NPOV policy cannot be superseded by other policies or guidelines, nor by editor consensus, and the editorial team enforces it.

As a result of the guideline, all Wikipedia material must be objective (neutral point of view). The Wikipedia team does not specify the file size for Wikipedia. Since it is the most extensive encyclopedia on the WWW, it is assumed that complete articles are large enough to be utilized for word embedding. It is reasonable to infer that Wikipedia is suitable for use as an objectivity dataset for the reasons above.

3.1.2 Subjectivity Corpus: Embedding Data

For subjectivity, the data have to be a non-factual, non-neutral point of view, or simply a human's opinion, emotion, and judgment. The subjectivity corpus, like the Wikipedia dataset, must also be reasonably large.

A suitable subjectivity corpus is the Shopee user review dataset. The Shopee customer review dataset is a collection of product reviews collected directly from the e-commerce sector. The dataset consists of real-world customer evaluations from several product categories.

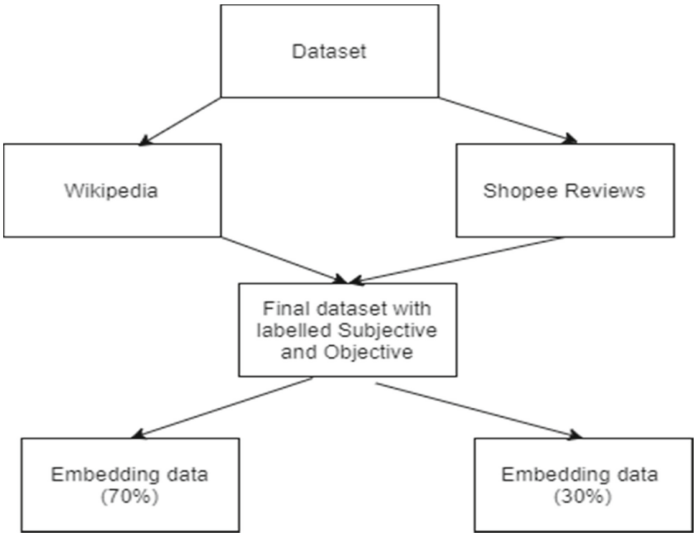

Fig. 7. Overview of the dataset preparation in phase one.

Shopee is a leading company in the digital economy. As a result, it is thought that the Shopee user review dataset is large enough to be utilized as the subjective corpus for embedding data. The product evaluations in the Shopee user review dataset are based on the purchasers' point of view, which comprises primarily personal ideas, judgments, and emotions. As a result of this critical takeaway, it is reasonable to use the Shopee user review dataset as a candidate for subjectivity corpus for embedding data.

3.1.3 Dataset Taxonomy

In summary, two primary datasets are being used in this research project: Wikipedia data and Shopee user reviews data. Both datasets will be merged and labeled as Objectivity and Subjectivity corpora. With 70% of the data being trained to create word vectors for embedding data, the as the test set. Figures 7 and 8 below show this research project's main dataset for phases one and two, respectively.

3.2 Data Preprocessing

3.2.1 Data Cleaning

In this part of the process, superfluous or unnecessary words are removed that are generally not essential or useful and may impede the process or give inaccurate results. Several data cleaning techniques are employed in stages:

(a) Remove punctuation and special characters
(b) Emoji cleaning
(c) Remove stop words

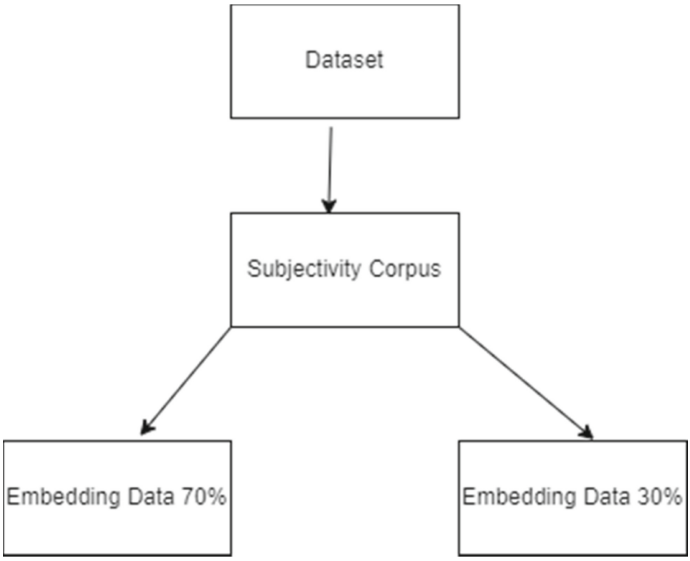

Fig. 8. Overview of the dataset preparation in phase two.

(d) Lowering case
(e) Lemmatization

In removing stop words, the aim is to remove common or meaningless words. For example, 'the', 'are', 'is' etc. The lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. To illustrate this point, the words 'cars', 'car's' are converted to 'car'.

3.2.2 Data Imbalance Treatment and Data Augmentation

A data imbalance issue arose in the subjectivity dataset, which is the Shopee dataset. Class imbalance is a situation that occurs when there is an unequal distribution of classes in a dataset, i.e. the number of data points in the positive class (majority class) is much more than that of the negative class (minority class). To address this, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to the dataset.

AUG-BERT, a contextualized word embedding approach from BERT's language model, is also applied to the Shopee dataset to increase the size of the corpus, through text augmentation. Both augmentation techniques seek to achieve a ratio of 50:50%.

3.3 Feature Extraction

Word2Vec, is used to train the word vectors which will be employed to classify subjectivity, objectivity, as well as the sentiment from the subjectivity corpus. Text features will be generated via Word2Vec for supervised machine learning methods.

In addition, TF-IDF is also considered for comparison. TF-IDF considers the frequency (TF) of a term and its inverse document frequency (IDF). Each word or phrase in the text has a TF and IDF score. The TF × IDF weight of a phrase is equal to the product of its TF and IDF ratings. The greater the TF × IDF score (weight), the more uncommon the word in a particular document is, and vice versa.

Finally, a BERT embedding is considered in this study. BERT is a computational paradigm that transforms words into numbers. As machine learning models require numbers as inputs rather than words, an algorithm that transforms words into numbers enables one to train machine learning models from original textual data.

3.4 Data Modelling

The objective of phase one is to determine if a sentence is objective or subjectivity from trained models of Word2Vec, TF-IDF, and BERT before sentiment analysis. Several classifiers are considered, namely:

(a) Random Forest
(b) Logistic Regression
(c) BERT
d) In addition, the models are also subject to:
(e) Parameter Tuning
(f) 70:30 split, 70% as training data, 30% as test data.

4 Results and Discussions

This research project's outcomes are divided into two phases – the first being subjectivity classification, from which the identified subjective statements are channeled to the second phase for sentiment analysis, as only subjective statements have elements of sentiment. Tables 5 and 6 show the results for subjectivity classification while Tables 8, 9, and 10 show the results for sentiment classification.

From Tables 5 and 6, the BERT embedding has better performance compared to the TF-IDF and Word2Vec embeddings in subjectivity classification. BERT embedding with the BERT classifier achieved 99.77% accuracy. Excluding the deep learning models from Table 9, the Word2Vec with Random Forest has the best performance, having an accuracy of 99.63%.

Tables 7 and 8 show the sentiment analysis classification results without data augmentation, while Tables 9 and 10 show sentiment classification results with data augmentation. The Word2Vec with Random Forest performed the best, achieving 83.80% without data augmentation. For data augmentation, the highest accuracy is 91.29% from TF-IDF with Random Forest.

5 Conclusions

Using corpora from Shopee Product Reviews and Wikipedia, word embeddings for subjectivity analysis were compared, namely Word2Vec, TF-IDF and BERT. Word embeddings from BERT performed the best for subjectivity analysis.

Table 5. Objectivity and subjectivity classification – accuracy

Objectivity and Subjectivity Classification				
Classifiers	Word Embedding			
	Word2Vec		TF-IDF	BERT
	Subjectivity Corpus	Objectivity Corpus		
Random Forest	99.63	97.71	99.42	98.23
Logistic Regression	99.46	97.26	99.62	99.18
BERT				**99.77**

Table 6. Objectivity and subjectivity classification – precision, recall and F1-score

Models	Metrics		
	Precision	Recall	F1-Score
Word2Vec Subjectivity + RF	99	99	99
Word2Vec Subjectivity + LR	100	100	100
Word2Vec Objectivity + RF	98	98	98
Word2Vec Objectivity + LR	97	97	97
TF-IDF + RF	99	99	99
TF-IDF + LR	100	100	100
BERT + RF	98	98	98
BERT + LR	99	99	99
BERT + BERT	100	100	100

Table 7. Sentiment analysis – accuracy (without data augmentation)

Classifiers	Word Embedding		
	Word2Vec	TF-IDF	BERT
Random Forest	**83.80**	83.08	78.12
Logistic Regression	79.75	80.35	75.01
BERT			82.71

In terms of sentiment analysis, using only subjectivity corpus, with the assumption that objectivity corpus does not contain elements of sentiment, TF-IDF with data augmentation, achieved the highest accuracy with 91.29%. The consideration of data augmentation provided improvement in the results as this increase the size of the corpus, and in turn affect the training for the better.

Table 8. Sentiment analysis – precision, recall and F1-score (without data augmentation)

Models	Metrics		
	Precision	Recall	F1-Score
Word2Vec Subjectivity + RF	**84**	**84**	**84**
Word2Vec Subjectivity + LR	79	79	79
TF-IDF + RF	83	83	83
TF-IDF + LR	80	80	80
BERT + RF	79	79	79
BERT + LR	75	75	75
BERT + BERT	83	83	83

Table 9. Sentiment analysis – accuracy (with data augmentation)

Classifiers	Word Embedding		
	Word2Vec	TF-IDF	BERT
Random Forest	87.79	**91.29**	79.01
Logistic Regression	79.52	81.38	74.63
BERT			91.15

Table 10. Sentiment analysis – precision, recall and F1-score (with data augmentation)

Models	Metrics		
	Precision	Recall	F1-Score
Word2Vec Subjectivity + RF	88	88	88
Word2Vec Subjectivity + LR	80	79	79
TF-IDF + RF	92	92	92
TF-IDF + LR	82	81	81
BERT + RF	80	79	79
BERT + LR	75	75	75
BERT + BERT	91	91	91

References

Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). The impact of features extraction on the sentiment analysis. *Procedia Computer Science*, *152*, 341–348.

Aung, K. Z., & Myo, N. N. (2017, May). Sentiment analysis of students' comment using lexicon based approach. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)* (Pp. 149–154). IEEE.

Ain, Q. T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., & Rehman, A. (2017). Sentiment analysis using deep learning techniques: a review. *International Journal of Advanced Computer Science and Applications*, *8*(6).

King J. (2017, August). English Wikipedia Articles 2017–08–20 SQLite. Retrieved 10 September, 2021 from https://www.kaggle.com/datasets/jkkphys/english-wikipedia-articles-20170820-sqlite.

Munikar, M., Shakya, S., & Shrestha, A. (2019, November). Fine-grained sentiment classification using BERT. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)* (Vol. 1, pp. 1–5). IEEE.

Poornima, A., & Priya, K. S. (2020, March). A comparative sentiment analysis of sentence embedding using machine learning techniques. In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 493–496). IEEE.

Soumya, S., & Pramod, K. V. (2020). Sentiment analysis of malayalam tweets using machine learning techniques. *ICT Express*, *6*(4), 300–305.

Suitnatsnoc L. (2020, August). Shopee Code League 2020 Data Science. Retrieved 10 September, 2021 from https://www.kaggle.com/datasets/davydev/shopee-code-league-20.