OPEN ACCESS
Indo-JC
Indonesia Journal of Computing
ISSN 2460-9056
socj.telkomuniversity.ac.id/indojc

Ind. Journal on Computing
Vol. 7, Issue. 2, August 2022. pp. 15-24
doi:10.34818/indojc.2022.7.2.638

Sentiment Analysis of University Social Media Using Support Vector Machine and Logistic Regression Methods

Fazainsyah Azka Wicaksono[#1], Ade Romadhony[#2], Hasmawati[#3]

# Faculty of Informatics, Telkom University
Jl. Telekomunikasi, Bandung, Jawa Barat, Indonesia

[1] fazainsyah@students.telkomuniversity.ac.id
[2] aderomadhony@telkomuniversity.ac.id
[3] hasmawati@telkomuniversity.ac.id

Abstract

Social media is one of the most powerful platforms for information sharing. Colleges or universities have official social media profiles to convey information about the campus and boost its branding and popularity. It is important for a university to comprehend its performance from the community's perspective, whether positive, negative, or indifferent toward the university. One solution is to examine the university's social media sentiment to establish the public's perception of the university. In this study, sentiment analysis was carried out on university social media based on public opinion or comments on every upload that the university had made to analyze and evaluate the performance of the university whether it was “Positive”, “Negative”, or “Neutral”. The extracted results can support the university's experience of criticism in measuring vital indicators. To classify posts on university Instagram, we use two methods: Support Vector Machine and Logistic Regression. The results suggest combining the Support Vector Machine approach with the TF-IDF feature yields the best F1-Score performance. In contrast, Logistic Regression with the FastText feature produces the worst performance of all models and feature extraction employed.

Keywords: Sentiment Analysis, Social Media, University, Support Vector Machine, Logistic Regression

Abstrak

Media sosial telah menjadi salah satu tempat untuk berbagi informasi yang sering kali digunakan oleh banyak masyarakat. Pada masa kini sudah lazim bagi perguruan tinggi atau universitas memiliki media sosial resmi untuk menyampaikan informasi seputar universitas dan meningkatkan *branding* atau popularitas universitas tersebut. Penting untuk sebuah universitas untuk mengetahui kinerja mereka berdasarkan pandangan publik, baik itu positif ataupun negatif terhadap universitas tersebut. Salah satu pendekatan untuk mengatasi hal ini adalah dengan melakukan analisis sentimen media sosial universitas untuk mengetahui opini publik terhadap universitas. Dalam penelitian ini, dilakukan analisis sentimen pada media sosial universitas berdasarkan opini atau komentar publik pada setiap unggahan yang telah dilakukan universitas untuk menganalisis sekaligus mengevaluasi kinerja dari universitas tersebut apakah “Positif”, “Negatif”, ataupun “Netral”. Hasil dari penelitian ini dapat mendukung pengalaman kritik universitas dalam mengukur indikator vital. Pada penelitian ini, metode yang digunakan untuk melakukan klasifikasi pada Instagram universitas adalah *Support Vector Machine* dan *Logistic Regression*, kemudian kami menerapkan dan melakukan analisis komparatif kinerja dari beberapa model tersebut. Hasil akhir menunjukkan bahwa penggunaan metode *Support Vector Machine* dengan fitur TF-IDF mendapatkan hasil performa *F1-Score* terbaik dan *Logistic Regression* dengan fitur *FastText* menghasilkan hasil performa paling terendah dari semua model dan ekstraksi fitur yang telah digunakan.

Received on April, 2022. Accepted on July, 2022

Kata Kunci: Analisis Sentimen, Media Sosial, Universitas, *Support Vector Machine, Logistic Regression*

I. Introduction

In the twenty-first century, the Internet has evolved into a platform for communicating knowledge that many people can rely on, and social media is one of these platforms. People have used social media to communicate information about their daily lives or help them with their work, particularly by sharing photos, videos, messages, and other content. Today, it is common for schools and colleges to have official social media profiles to disseminate information about the institution and improve brand awareness or popularity [1]. The outcomes of each post or upload issued by the institution include public comments that may positively or negatively influence the public's impression of the university.

Sentiment analysis, often known as opinion mining, categorizes the opinions expressed in text documents (positive or negative). With the advancement of Web 3.0 and the growing popularity of social media, there is an abundance of user-generated information on products, events, and services. Sentiment analysis is a useful technique for mining user-generated data patterns since it is often utilized for real-world situations. Research related to sentiment analysis has been carried out by various researchers such as sentiment analysis for product reviews (X. Fang and J. Zhan, 2015) [2], restaurant reviews (H. Kang et al., 2012) [3], microblogs, and social media posts twitter (F. Neri et al., 2012) [4]. Attitudes in textual data have been analyzed at different levels, including aspect, sentence, and document levels. Document-level sentiment analysis seeks to analyze the entire document. Sentiment analysis at the sentence level determines if a sentence indicates a positive or negative opinion.

In contrast, aspect-based sentiment analysis determines sentiment at a finer level, such as towards an entity. Sentiment analysis can be used to analyze comments made by students, teachers, or members of the general public in the form of opinions or college components. It is possible to do sentiment analysis to determine how each community perceives the college or university, and sentiment analysis may also help universities evaluate their performance. This sentiment survey yielded three results: “Positive,” “Negative,” and “Neutral [5].”

The main contribution of this research is to conduct sentiment analysis on university social media using the comments on each upload provided by the college or university to analyze and evaluate the performance of the university whether it was “Positive”, “Negative”, or “Neutral”. The extracted results can support the university's experience of criticism in measuring vital indicators to evaluate the university and establish the public's perception of the university. The dataset used in this study is one with labeled data that has been manually labeled and applied to the dataset. The methods that will be used in this study are Logistic Regression (LR) and Support Vector Machine (SVM). The selection of Logistic Regression and Support Vector Machine methods' is based on the efficiency and proven performance in conducting sentiment analysis on a variety of problems [6]. Giatsoglou et al. (2017) found that the SVM model achieves the best results in terms of efficiency in accuracy and process times [7]. H. Hamdan et al. (2015) said that using the LR model can compare with other models such as SVM in sentiment polarity [8]. Based on the two approaches, comparative analysis in the form of performance benchmarks for the results of the algorithms used will be provided. One way to measure university performance is to look at public opinion or sentiment which can be seen through comments made on its social media platforms.

II. Literature Review

A. Sentiment Analysis

Sentiment analysis, also known as opinion mining, is a field of study that examines people's opinions, sentiments, evaluations, reviews, attitudes, and feelings toward a sure thing, such as products, services, organizations, individuals, problems, events, subjects, and traits [9]. The three levels of sentiment analysis are document level, sentence level, and entity and aspect level. Several previous studies have used sentiment analysis on a college's Twitter social media utilizing Support Vector Machine (SVM) and Naive Bayes to collect data based on tweets related to their university [5]. On Instagram, sentiment analysis data may be evaluated by looking at the number of sentiment-related comments on a specific upload or all comments on an account.

1) Document Level: At this point, it is assessed whether the overall document displays a positive or negative perspective. When community member comments on University social media, the system determines if the statement reflects a “positive” or “negative” attitude toward the university, this method is also known as Document Level Sentiment Classification. At the document level, each document expresses an opinion about a single entity (for example, one college). As a result, it does not apply to documents that evaluate or compare multiple entities.

2) Sentence Level: The challenges for this level are to construct sentences and determine whether each sentence represents a “positive,” “negative,” or “neutral” sentiment. Typically, neutral means that the sentence contains no opinion. The Sentence Level is closely related to subjectivity classification, which distinguishes between sentences that reveal factual information (referred to as objective sentences) and corrections that express subjective views and opinions (referred to as subjective sentences) (called subjective sentences). A “positive” opinion includes the phrase “this school is really good.”

3) Entity and Aspect Level: Individuals’ genuine likes and dislikes were not revealed by either document-level or sentence-level studies. At the aspect level, a finer analysis is carried out. The aspect level investigates opinions rather than language constructions (documents, paragraphs, phrases, or clauses). It is based on the idea that an opinion is made up of a sentiment (“positive” or “negative”) and a target (opinion).

B. Logistic Regression

Logistic Regression is the most recent version of Linear Regression since it allows for analyzing binary outcomes with two mutually exclusive levels. Logistic Regression allows for using continuous or categorical predictors and changing various variables. Previous studies regarding Logistic Regression in carrying out sentiment analysis, opinion target extraction by H. Hamdan et al. [8], and sentiment analysis by S. Sazzed and S. Jayarathna using Logistic Regression and Support Vector Machine as their training model [6].

C. Support Vector Machine

A Support Vector Machine (SVM) is a machine learning algorithm based on learning theory. These methods are similar to Artificial Neural Network (ANN), which is used for estimation and linear data separation. SVM, unlike regression methods, does not often require variable interaction. These, like ANNs, should be able to hold many phases of pre-processing data, require non-lost numerical input, and are frequently robust to noise and outliers [10]. Then there is some previous study on performance using the Support Vector Machine for sentiment analysis [11] and social media sentiment analysis where A. Abdelrazeq et al. using SVM as one of their model because Support Vector Machine classifier can be applied to process on a large data [5].

D. F1-Score

F1-Score is a technique that averages precision and recall. F1-Score is derived from calculations based on precision and recall since it assigns equal weight to both factors. Due to classification issues, accuracy may not be appropriate because it does not account for the possibility of tuples or features belonging to multiple classes; consequently, Linear Regression or F-Measure is more appropriate [12]. Following is the calculation for Linear Regression:

$$P = \frac{TP}{TP + FP} \tag{1}$$

$$R = \frac{TP}{TP + FP} = \frac{TP}{P} \quad (2)$$

$$F_1 = \frac{2PR}{P + R} \quad (3)$$

The following is a definition of the variables in the preceding F1-Score (F_1) formula [12], as you can see in table I:

TABLE I
Definition Variable for Precision (P), Recall (R), and F1-Score (F_1) Formulas

Variable	Symbol	Definition
Precision	P	A precision metric indicating what proportion of tuples is genuinely labeled positive.
Recall	R	The measure of completeness or the proportion of positive tuples
True Positives	TP	The classifier correctly categorizes them as positive tuples.
True Negatives	TN	The classifier correctly categorizes them as negative tuples.
False Positives	FP	The classifier incorrectly categorizes them as positive tuples. If the actual class is no and the predicted class is yes.
False Negatives	FN	The classifier incorrectly categorizes them as negative tuples. If the actual class is yes and the predicted class is no.

E. TF-IDF

Term Frequency Inverse Document Frequency (TF-IDF) is a statistical technique that allows weights to be assigned to each document term (or word). The number of times a word appears in a document is represented by Term Frequency (TF), whereas the number of times a word appears in the document is represented by Inverse Document Frequency (IDF). Natural Language Processing (NLP), information retrieval, and text mining are all applications of this method. The method can determine the importance, magnitude, and weight of a term's (or word's) meaning in a collection of documents. The importance of text grows in direct proportion to its frequency in the document [13].

F. Bag-of-Words

The Bag-of-Words (BOW) model is a straightforward representation used in Natural Language Processing (NLP) and Information Retrieval (IR). This model represents text as an irregular word collection, disregarding grammar and word order. In-text categorization, a document's words are weighted according to their frequency inside the document and across documents. The BOW is composed of words of greater weight [14].

G. Word Embeddings

Word representation is a crucial aspect of natural language processing. The usual method of expressing words as separate and independent symbols is inadequate and insufficiently generic for numerous activities. Recently, it has been proposed that words be represented as dense vectors obtained from various training methodologies influenced by neural network language modeling. These representations, known as “neural embeddings” or “word embeddings,” have been demonstrated to be effective for several tasks [15]. Tomáš Mikolov introduced and implemented the skip-gram model with negative sampling (SKIPGRAM) in his software word2vec as one of the advanced word embedding approaches [16].

In addition, FastText is one of the excellent feature extraction tools that can perform the task of detecting word difficulty because it can use word morphology information to generalize it. FastText was trained using CBOW by combining weights that depend on position, and Skipgram uses n-gram characters associated with vector representations [17]. Word embeddings provide meaningful word representations and are highly efficient for training, working online, and scaling to the massive corpus (billions of words) and huge word and context vocabulary [15].

III. Research Method

A. Dataset

This study uses Telkom University's official Instagram social media from January 10, 2021, to December 31, 2021, which 251 uploads. Post_username, post_time, comment_time, post_like, post_view, comment_user, and comment_txt are fields in the raw data. Each existing piece of data is labeled with a description of how it makes the person feel. The dataset only has the comment feature, where every emoji in the comments has been converted to text, as in Table II. Each comment is then labeled with positive, negative, and neutral. Table III shows examples of comments that have been labeled.

TABLE II
An example of converting emoji to text.

No.	Original Comment	Conversion Result
1.	Gilaaa Kereenn 🔥🔥👏	Gilaaa kereenn fire fire clapping hands
2.	Selamat Ulang tahun Prof ❤️	Selamat Ulang tahun Prof red heart
3.	pengen ke sini 😍	pengen ke sini smiling face with heart eyes
4.	Salut sama telkom 👍👍	Salut sama telkom thumbs up thumbs up
5.	Woww 🙌	Woww raising hand

TABLE III
Example of a labeled comment

No.	Comment	Label
1.	Gilaaa kereenn fire fire clapping hands	positive
2.	Selalu no 1 aamin..	positive
3.	miris kampusku sekarang ... untung udh minggat wkwkwkwkwk	negative
4.	dih kocak bgt 200rbmah dipake bayar wifi aja kurang	negative
5.	Min kalo mahasiswa telkom untuk vaksin ke dua bisa ga kira kira?	neutral
6.	Hai min, saya dm dari tgl 30 juli untuk urus ijazah, tp blm dibalas. Apa bisa dibantu?	neutral

B. Pre-Processing

Before proceeding to the training and testing phases, the comment text data from each upload must be pre-processed. The following techniques will be employed:

1) **Capitalization/Case Folding***:* Change all existing uppercase letters to lowercase.

2) **Strip Punctuation***:* Remove all punctuation from a sentence, including “,”, “.”, “!”, etc.

3) **Tokenization***:* Divide sentences into word units (tokens).

4) **Stemming***:* To simplify each word to its most fundamental form.

5) **Undersampling***:* To delete data randomly from the majority of data, which makes the dataset unbalanced. This is an example of undersampling when there isn't as much “negative” data as “positive” data. The total number of “positive” data will follow the total number of “negative” data.

C. Data Split

The data will be divided into two distinct groups that will not overlap for the training and testing of the model. In this study, the dataset will be divided into 70% for training and 30% for testing.

D. Feature Extraction

Because the machine learning model will not accept input in the form of strings, the feature extraction method must first be performed on the metadata features present within the dataset. Each word in the text will be assigned a weight and employed as a vector depending on the data in the document. The TF-IDF method [13], the Bag-of-Words (BOW) method [14], and the FastText method are used to compare feature extraction methods [18]. FastText is a word embedding method that employs a consistent methodology to express words as distinct and distinct symbols. It is insufficient for many professions and is not generic enough [15].

E. Model Training

The two methods that will be used in this inquiry are Logistic Regression (LR) [8] and Support Vector Machine (SVM) [11]. Both models are trained using the same data from the training set. The parameter setups are different based on the model used. Because there is a difference in the amount of data on the "positive" label and the "negative" label, the data used in the training model is the result of undersampling based on the amount of data on the "negative" label. The trained train data is the result of the training model and will be used to categorize the test data once it has been trained.

F. Model Testing

Following the training phase, the models will proceed to the testing phase. They will be evaluated using test data comprised of information the model has never seen before. The data will go through the same pre-processing stage as the training data, and then the two trained models will generate label predictions based on the data in the test data. The result of the testing phase is the prediction label for each model used on test data.

G. Evaluation Metrics

The performance of each model is evaluated in terms of precision, recall, and F1-Score on "positive," "neutral," and "negative" sentiments. The term "precision" refers to the percentage of correct entities found by the system, as described in formula 1, whereas "recall" refers to the percentage of entities discovered by the system, as described in formula 2. The "F1-Score" is calculated by adding together the "precision" and "recall", as mentioned in formula 3 [12]. The prediction is correct if the entity extracted by the system matches the entity in the data.

IV. Results And Discussion

H. Testing Results and Analysis

The dataset used was crawling from Telkom University's Instagram post from January 10, 2021, to December 31, 2021. In this dataset, there are 251 posts with a total of 1864 comments. The comments in the dataset have been labeled into 3 classes, "positive", "negative", and "neutral". The distribution of the dataset for each sentiment can be seen in Table IV.

TABLE IV
Distribution of each sentiment before augmented data

No.	Data Sentiment	Quantity
1.	positive	809
2.	negative	91
3.	neutral	946

Based on the experiments conducted using the dataset in Table IV, the model performance was obtained as shown in Table V.

TABLE V
THE RESULT FROM THE MODEL WITHOUT AUGMENTED DATA

Model	Feature Extraction	Sentiment	Precision	Recall	F1-Score	F1-Score(Avg)
Logistic Regression	TF-IDF	positive	89%	79%	84%	63.3%
		neutral	80%	67%	73%	
		negative	21%	75%	33%	
	BOW	positive	88%	78%	83%	**66.3%**
		neutral	80%	74%	76%	
		negative	27%	75%	40%	
	FastText	positive	70%	76%	73%	48.6%
		neutral	75%	37%	50%	
		negative	14%	75%	40%	
Support Vector Machine	**TF-IDF**	positive	85%	79%	82%	**62.3%**
		neutral	79%	65%	71%	
		negative	22%	75%	34%	
	BOW	positive	79%	78%	79%	60.6%
		neutral	78%	60%	68%	
		negative	22%	79%	35%	
	FastText	positive	73%	88%	80%	51.3%
		neutral	82%	34%	48%	
		negative	15%	82%	26%	

Table V reveals that testing for each model with different feature extraction produces different results. The Logistics Regression model with Bag-of-Words (BOW) as feature extraction produces the best results, with an average F1-score of 66.3%. The best results from the Support Vector Machine model are obtained when using TF-IDF as feature extraction, with an average F1-Score of 62.3 %. Based on the three feature extractions used, the prediction results from FastText are smaller than the others. This is due to insufficient data, comments on the data have slang words or abbreviations that can affect the performance of the model when using FastText. Since FastText itself uses a language corpus where the model will look for words that have the same meaning in every n-gram. However, these results indicate an underwhelming performance. Particularly in F1-Score results with a “negative” sentiment. Therefore, based on the result obtained, we add the amount of data with negative sentiments to evaluate the performance of the model. In this case, we add 136 augmented data with “negative” sentiments. The distribution of the amount of data on each sentiment after the addition of negative sentiment data can be seen in Table VI.

TABLE VI
DISTRIBUTION OF EACH SENTIMENT AFTER AUGMENTED DATA

No.	Data Sentiment	Quantity
1.	positive	809
2.	negative	245
3.	neutral	946

After adding negative sentiment data, we evaluate the performance of both models. The results of the model's performance after the addition of negative sentiment data can be seen in Table VII.

TABLE VII
The result from the model with augmented data

Model	Feature Extraction	Sentiment	Precision	Recall	F1-Score	F1-Score(Avg)
Logistic Regression	TF-IDF	positive	87%	76%	81%	73.3%
		neutral	77%	75%	76%	
		negative	52%	79%	63%	
	BOW	positive	87%	79%	83%	**74%**
		neutral	78%	77%	78%	
		negative	53%	72%	61%	
	FastText	positive	79%	77%	78%	66.6%
		neutral	78%	60%	68%	
		negative	41%	80%	54%	
Support Vector Machine	**TF-IDF**	positive	86%	82%	84%	**74.6%**
		neutral	81%	74%	77%	
		negative	52%	79%	63%	
	BOW	positive	78%	82%	80%	69.6%
		neutral	79%	62%	69%	
		negative	48%	80%	60%	
	FastText	positive	80%	85%	82%	71.3%
		neutral	83%	67%	74%	
		negative	49%	75%	59%	

Table VII shows the performance of the method after the augmentation of data. When using Logistic Regression, the best results are obtained by using Bag-of-Word (BOW) with an average F1-Score of 74%. The use of the TF-IDF produces the best results for the Support Vector Machine model. The results obtained are an average F1-Score of 74.6%.

With the addition of augmented data, the overall performance of the model has improved especially for data with “negative” sentiments ranging from 21% to 33%. The largest performance increase was achieved by the Support Vector Machine model using FastText, with an increase of 33% as can be seen in tables V and VII. This result achieved was also affecting the average F1-Score in each model which increased from 7.6% to 20%. The largest significant rise was achieved by Support Vector Machine with FastText, which increased by 20%. From these results, it can be seen that the model that uses FastText produces a significant improvement compared to the model with TF-IDF or Bag-of-Word (BOW). This is due to the addition of augmented data helps FastText in studying the data. Although the model is quite solid, the model works are influenced by the data with “negative” sentiments. This could be due to the lack of data in the dataset, the amount of data with “negative” sentiment is less than the “positive” and “neutral” sentiment. Therefore, an evaluation of the addition of augmented data is needed. After under-sampling the dataset, the results obtained are more stable because the comparison between the three sentiment classes is more balanced. However, the lack of available data affects the prediction results of the Support Vector Machine and Logistic Regression methods so that it is not optimal because the model can only learn based on the data it has.

V. Conclusion

According to the findings of this study, the Logistic Regression and Support Vector Machine models, along with the use of the TD-IDF, Bag-of-Words, and FastText features, can be used to determine whether a university's performance is good or bad based on public sentiment on university social media. The best results are obtained when working on models with augmented data. Based on the outcome, Support Vector Machine can produce better results than Logistic Regression. By using the TF-IDF, the Support Vector Machine gets an average F1-Score of 74.6%. Meanwhile, Logistic Regression works best when using the Bag-of-Words with an average F1-Score of 74%. The biggest improvement after adding augmented data to each model is by using FastText as feature extraction, it increases higher than other feature extractions, which is an increase of 20%.

The results also show that by under-sampling based on “negative” sentiment, the performance obtained is more stable. The weakness of this model is that the distribution for each sentiment is not even, especially for the “negative” sentiment which is less compared to the “positive” and “neutral” sentiment. For further research, we suggest adding a balanced number of sentiments for each “negative”, “positive”, and “neutral” sentiment.

References

[1] R. Rutter, S. Roper, and F. Lettice, “Social media interaction, the university brand and recruitment performance,” *J. Bus. Res.*, vol. 69, no. 8, pp. 3096–3104, 2016, doi: https://doi.org/10.1016/j.jbusres.2016.01.025.

[2] X. Fang and J. Zhan, “Sentiment analysis using product review data,” *J. Big Data*, vol. 2, no. 1, p. 5, Jun. 2015, doi: 10.1186/s40537-015-0015-2.

[3] H. Kang, S. J. Yoo, and D. Han, “Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews,” *Expert Syst. Appl.*, vol. 39, no. 5, pp. 6000–6010, 2012, doi: https://doi.org/10.1016/j.eswa.2011.11.107.

[4] F. Neri, C. Aliprandi, F. Capeci, M. Cuadros, and T. By, “Sentiment Analysis on Social Media,” Aug. 2012.

[5] A. Abdelrazeq, D. Janßen, C. Tummel, S. Jeschke, and A. Richert, “Sentiment analysis of social media for evaluating universities,” in *Automation, Communication and Cybernetics in Science and Engineering 2015/2016*, Springer, 2016, pp. 233–251.

[6] S. Sazzed and S. Jayarathna, “SSentiA: A Self-supervised Sentiment Analyzer for classification from unlabeled data,” *Mach. Learn. with Appl.*, vol. 4, p. 100026, Jun. 2021, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666827021000074

[7] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas, “Sentiment analysis leveraging emotions and word embeddings,” *Expert Syst. Appl.*, vol. 69, pp. 214–224, 2017, doi: https://doi.org/10.1016/j.eswa.2016.10.043.

[8] H. Hamdan, P. Bellot, and F. Bechet, “Lsislif: Crf and logistic regression for opinion target extraction and sentiment polarity analysis,” in *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, 2015, pp. 753–758.

[9] M. Adnan, R. Sarno, and K. R. Sungkono, “Sentiment Analysis of Restaurant Review with Classification Approach in the Decision Tree-J48 Algorithm,” *2019 Int. Semin. Appl. Technol. Inf. Commun.*, pp. 121–126, 2019.

[10] S. Garc\'\ia, J. Luengo, and F. Herrera, *Data preprocessing in data mining*, vol. 72. Springer, 2015.

[11] N. Zainuddin and A. Selamat, “Sentiment analysis using support vector machine,” in *2014 international conference on computer, communications, and control technology (I4CT)*, 2014, pp. 333–337.

[12] J. Han, M. Kamber, and J. Pei, “8 - Classification: Basic Concepts,” in *Data Mining (Third Edition)*, Boston: Morgan Kaufmann, 2012, pp. 327–391. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123814791000083

[13] B. Trstenjak, S. Mikac, and D. Donko, “KNN with TF-IDF based framework for text categorization,” *Procedia Eng.*, vol. 69, pp. 1356–1364, 2014.

[14] K. Soumya George and S. Joseph, “Text classification by augmenting bag of words (BOW) representation with co-occurrence feature,” *IOSR J. Comput. Eng*, vol. 16, no. 1, pp. 34–38, 2014.

[15] O. Levy and Y. Goldberg, “Dependency-based word embeddings,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2014, pp. 302–308.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv Prepr. arXiv1301.3781*, 2013.

[17] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 135–146, 2017.

[18] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, “Learning Word Vectors for 157 Languages,” 2018.