p-ISSN: 2252-3006
e-ISSN: 2685-2411

Text Classification System Using Text Mining with XGBoost Method

Ni Kadek Dwi Rusjayanthi[a1], Anak Agung Kompiang Oka Sudana[a2], I Nyoman Prayana Trisna[a3]
[a]Department of Information Technology, Udayana University, Badung, Indonesia
e-mail: [1] dwi.rusjayanthi@unud.ac.id, [2]agungokas@unud.ac.id, [3]prayana.trisna@unud.ac.id

Abstrak

Data berukuran besar di masa kini dapat dimanfaatkan untuk analisis, sehingga dapat diperoleh pengetahuan penting/bermanfaat pada berbagai domain. Analisis teks dapat dilakukan memanfaatkan text mining menggunakan metode komputasi sehingga ekstraksi pengetahuan dapat dilakukan pada data teks berukuran besar, termasuk pemrosesan terkait data teks yang tidak terstruktur, yang ditulis dalam bahasa alami. Klasifikasi pada text mining adalah salah satu tipe pekerjaan dengan proses pencarian sekumpulan model atau fungsi yang menggambarkan dan membedakan kelas data text dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas dari suatu objek (data text) yang belum diketahui kelasnya. Text mining dilakukan pada penelitian ini untuk analisis data teks melalui Sistem Klasifikasi Teks menggunakan salah satu metode klasifikasi yaitu Metode XGBoost (eXtreme Gradient Boosting). Sistem klasifikasi teks dikembangkan untuk mengklasifikasikan teks berupa artikel. Akurasi tertinggi yang diperoleh dari pengujian yaitu sebesar 77%, dengan presisi sebesar 81% dan recall sebesar 77%.

Kata kunci: *text mining, data teks, klasifikasi, Metode XGBoost*

Abstract

Large data nowadays can be used for analysis; thus, it can obtain important/valuable knowledge in various domains. Text analysis can be carried out by utilizing text mining using computational methods so that knowledge extraction can be carried out on large text data, including processing related to unstructured text data, which is written in natural language. Classification in text mining is a type of work with the searching process for a set of models or functions that describe and differentiate text data classes with the aim that the model can be used to predict the class of an object (text data) whose class is unknown. Text mining was carried out in this research to analyze text data through the Text Classification System using a classification method, namely the XGBoost (eXtreme Gradient Boosting) Method. A text classification system was developed to classify text in the form of articles. The highest accuracy obtained from the test is 77%, with a precision of 81% and a recall of 77%.

Keywords: *text mining, text data, classification, XGBoost Method*

1. Introduction

The availability of massive data nowadays can be used for analysis; thus, it can obtain important/useful knowledge in various domains. Text mining can be used for text analysis using computational methods so that knowledge extraction can be carried out on large text data, including processing related to unstructured text data written in natural language.

Text mining is a process of extracting information, where users interact with documents using analytical tools in the form of data mining components, including clustering components. Text mining adopts various techniques from other fields, such as data mining, information retrieval, machine learning, statistics and mathematics, linguistics, natural language processing (NLP), and visualization. Activities related to research for text mining include text extraction and storage, preprocessing, statistical data collection, indexing, and content analysis [1]. Text mining tasks include text categorization, text clustering, concept/entity extraction, sentiment analysis, document summarization, and entity-relation modeling. Text mining is used to process unstructured data, in contrast to data mining which is used to process structured data [2].

Classification is the searching process or a set of models or functions that describe and differentiate data classes. The goal of classification is that the model can be used to predict the class of an object whose class is unknown [3]. Classification is often referred to as the supervised method because it uses previously labeled data as examples of correct data. Text classification is a classification process on textual data that has been carried out in several studies using various methods. Terms often used in text data include documents, words, phrases, corpus, and lexicon. Documents are sequences of words and punctuation marks that follow the grammatical rules of the language. Sentences, paragraphs, sections, chapters, books, web pages, emails, and more are some instances of the documents in this context. A term is usually a word but can also be a word or phrase pair. The corpus is a collection of documents, while the lexicon is a set of all the unique words in the corpus. Popular methods used for classification in the text include Nearest Neighbor (NN), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), and Neural Networks method [4].

Text classification utilizing embedded representation of words and word sense has been studied and produced a stable classification process, especially in classifications with complex semantics [5]. Local features of phrases and global sentence semantics have been used for text classification using the AC-BiLSTM (attention-based bidirectional long short-term memory with convolution layer) method [6]. In addition, word embedding has also been used for text classification using various classification methods [7]. However, the method commonly used is TFIDF. TFIDF is a method that integrates term frequency and inverse document frequency. Term frequency calculated by the i-th term is the frequency of the t-th term appearing in the d-th document. Inverse Document Frequency helps reduce the influence of common words in the corpus [8].

Text mining in this research analyzed text data through a text classification system using classification techniques. The classification method used was the XGBoost (eXtreme Gradient Boosting) method, one of the tree-based machine learning methods that utilize tree-boosting techniques [9]. This study employed XGBoost because it enabled resource optimization through cache access patterns, data compression, and sharding. The data used in this research were Indonesian language article data. TFIDF (Term Frequency Inverse Document Frequency) for feature extraction, ANOVA (Analysis of Variance) for feature selection, and PCA (Principal Component Analysis) for feature dimension reduction were other techniques applied in this work. The XGBoost method has been studied in several studies related to classification, including building a milk source classification model (dairy farming) [10], diabetes prediction [11], traffic accidents prediction [12], gourami supply estimation [13], and landslide hazard mapping [14]. The utilization of the XGBoost method for text mining has been carried out in several studies, including the hybrid model development for Ukrainian language sentiment analysis [15], integrated technology analysis of patent data [16], classification of injury rates based on accident narrative data [17], and classification of proactive personality in social media users [18].

2. Research Method / Proposed Method

System development was carried out in two main stages: the training and the testing stage. System overview can be seen in Figure 1.

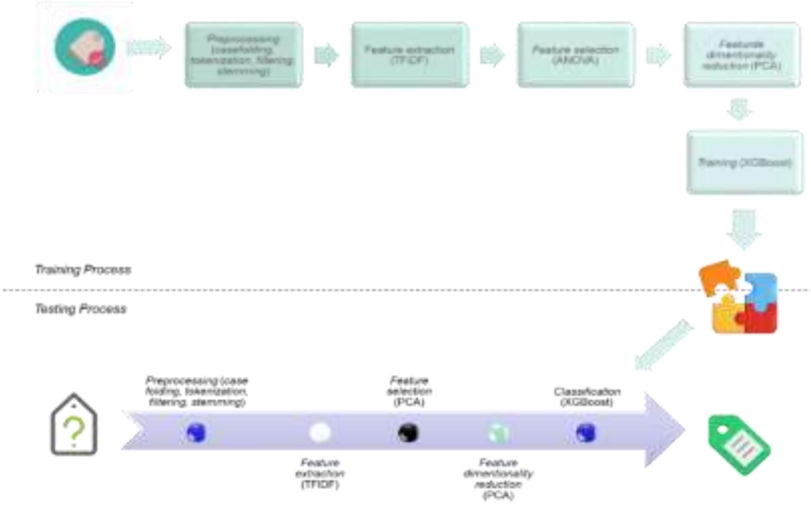

Preprocessing (casefolding, tokenization, filtering, stemming)
Feature extraction (TFIDF)
Feature selection (ANOVA)
Featurde dimentionality reduction (PCA)
Training (XGBoost)
Training Process
Testing Process
Preprocessing (case folding, tokenization, filtering, stemming)
Feature extraction (TFIDF)
Feature selection (PCA)
Feature dimentionality reduction (PCA)
Classification (XGBoost)

Figure 1 System overview

The training stage aimed to build a model, and the testing stage aimed to test system performance through a model formed using the classification method, namely XGBoost, during the training stage.

Articles classification based on Figure 2 began with document preprocessing. Preprocessing aimed to prepare text into data that could be processed at the next stage (training). The process carried out at the training stage includes (1) Case folding or changing all the letters in the article to lowercase; (2) Tokenization or the stage where a collection of characters in a text that has gone through a case folding process was broken down into units of words (tokens) with the first step, namely dividing the document into smaller parts, namely paragraphs and then sentences; (3) Filtering or the stage of taking important/related words by removing non-alphabetic characters and stopwords or meaningless words/related to determining the topic, as well as removing characters other than letters such as numbers, punctuation, whitespace or blank characters, while filtering in research was based on Sastrawi Algorithm and; (4) Stemming was the stage of mapping and decomposing the form of a word into its basic word form. This process also used Sastrawi Algorithm.

Furthermore, documents that have gone through preprocessing were subjected to TFIDF weighting or the term weighting process. Feature selection was used in feature extraction results to reduce feature size by selecting more relevant features. The feature selection method used was ANOVA (Analysis of Variance). Feature dimension reduction was applied to the feature from feature selection results to reduce feature dimensions. The method used for feature dimension reduction was PCA (Principal Component Analysis). The training was carried out using the XGBoost Method, which was applied to the TFIDF-weighted features of the training data document. The testing was carried out on test data (features) using the model produced during training as explanation of the stages of research that illustrates the logical sequence to get research output in line with expectations.

The text data used in this research were Indonesian articles obtained from the news site www.cnnindonesia.com. The data used consisted of five article topics: Economy, Sports, Entertainment, Lifestyle, and Technology. Each article topic consisted of 20 articles, so the total amount of data used was 100 article data.

3. Literature Study

3.1. Classification

Classification is a technique in Data Mining that is used to extract patterns/knowledge from text in Text Mining. The purpose/function of classification is knowledge extraction or model building to predict previously unknown class/category data [3]. The model is formed through the training phase, while the accuracy or performance of the model is obtained through the testing phase. Classification is included in the supervised learning category because the data used is labeled data or with the class column.

3.2. XGBoost (eXtreme Gradient Boosting) Method

XGBoost is a tree-based machine learning method which utilizes tree boosting techniques [9]. The utilization of cache access patterns, data compression, as well as sharding in XGBoost are the main components that can support more optimal use of resources. The tree boosting technique implemented in XGBoost is scalable, which is effective for preventing overfitting [19].

3.3. TF-IDF

TFIDF is a method/technique of word weighting using term frequency ($\mathrm{tf}_{t,d}$), and inverse document frequency (idf_t). The term frequency is obtained by calculating the frequency of a term/word (*t*) contained in document (*d*), and the inverse document frequency (idf_t) is obtained by the logarithmic ratio between the number of documents in the corpus (*N*) and the number of documents that have the term *t* (df_t). The inverse document frequency is useful for reducing the influence of common words on the corpus [8] , obtained by calculating using Equation (1), while the TFIDF weight is obtained using Equation (2).

$$\mathrm{idf}_t = \log_{10} N/\mathrm{df}_t \tag{1}$$

$$\mathrm{w}_{t,d} = (1 + \log \mathrm{tf}_{t,d}) \times \log \ N/\mathrm{df}_t \tag{2}$$

The TFIDF weighting function is to obtain values that can be used to represent documents of the training data.

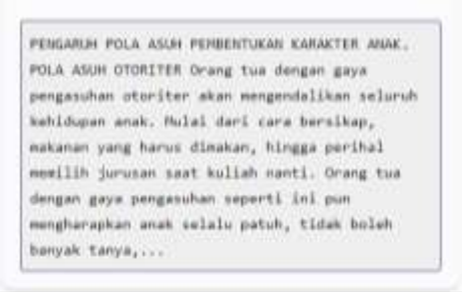
3.4. Feature Selection and Dimentionality Reduction

High-dimensional datasets with large feature sizes can cause various obstacles in the learning process in machine learning. These obstacles include increasing the dimensions of the search space and data preparation for the learning process, as well as increasing computational complexity [20]. Feature selection is a process of selecting attributes that are considered relevant in the machine learning process, including text mining. Reducing feature size is useful for saving training time and reducing model complexity, and can even support model performance. One of the feature selection methods is Anova (Analysis of Variance) method which utilizes variations between the average features and data attributes from various classes/groups. Dimensionality reduction is used to reduce feature size (dimensions). The method used in this study is PCA (Principal Component Analysis) which is obtained based on the principal data components with the highest value variations.

4. Result and Discussion

4.1. Preprocessing

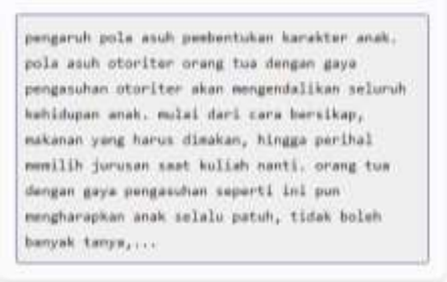
Preprocessing was conducted for training and testing data. The preprocessing stages consisted of case folding, tokenization, filtering, and stemming. An example of one of the initial article data before the preprocessing stage is shown in Figure 2.



```
PENGARUH POLA ASUH PEMBENTUKAN KARAKTER ANAK.
POLA ASUH OTORITER Orang tua dengan gaya
pengasuhan otoriter akan mengendalikan seluruh
kehidupan anak. Mulai dari cara bersikap,
makanan yang harus dimakan, hingga perihal
memilih jurusan saat kuliah nanti. Orang tua
dengan gaya pengasuhan seperti ini pun
mengharapkan anak selalu patuh, tidak boleh
banyak tanya,...
```

Figure 2 Example of raw text data

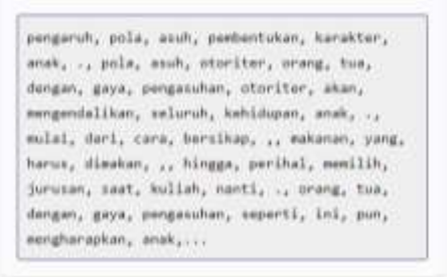
Case folding was applied to change text data to lowercase. The resulting data from the case folding process is shown in Figure 3, where words that previously consisted of uppercase letters have changed to lowercase letters.



```
pengaruh pola asuh pembentukan karakter anak.
pola asuh otoriter orang tua dengan gaya
pengasuhan otoriter akan mengendalikan seluruh
kehidupan anak. mulai dari cara bersikap,
makanan yang harus dimakan, hingga perihal
memilih jurusan saat kuliah nanti. orang tua
dengan gaya pengasuhan seperti ini pun
mengharapkan anak selalu patuh, tidak boleh
banyak tanya,...
```

Figure 3 The results of the case folding process

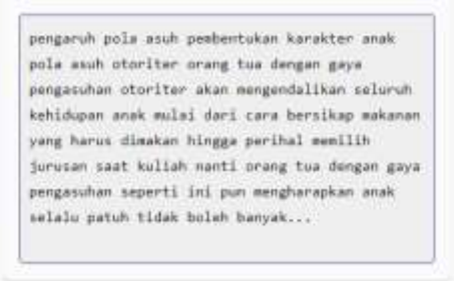
The results of the tokenization process are shown in Figure 4, where the formed data has been separated into units of words/tokens. The data, which originally consisted of paragraphs, was formed into smaller parts, namely sentences. From sentences, the data was formed into even smaller parts in the form of tokens/words.



```
pengaruh, pola, asuh, pembentukan, karakter,
anak, ., pola, asuh, otoriter, orang, tua,
dengan, gaya, pengasuhan, otoriter, akan,
mengendalikan, seluruh, kehidupan, anak, .,
mulai, dari, cara, bersikap, ,, makanan, yang,
harus, dimakan, ,, hingga, perihal, memilih,
jurusan, saat, kuliah, nanti, ., orang, tua,
dengan, gaya, pengasuhan, seperti, ini, pun,
mengharapkan, anak,...
```

Figure 4 The results of the tokenization process

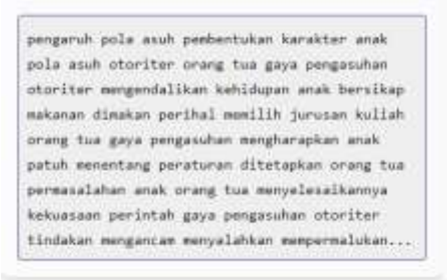
The resulting data from the filtering process is shown in Figure 5. There was an omission of non-alphabetical characters, such as numbers and punctuation marks, so the data only consisted of the letters a to z.



```
pengaruh pola asuh pembentukan karakter anak
pola asuh otoriter orang tua dengan gaya
pengasuhan otoriter akan mengendalikan seluruh
kehidupan anak mulai dari cara bersikap makanan
yang harus dimakan hingga perihal memilih
jurusan saat kuliah nanti orang tua dengan gaya
pengasuhan seperti ini pun mengharapkan anak
selalu patuh tidak boleh banyak...
```

Figure 5 The omission of non-alphabetical characters

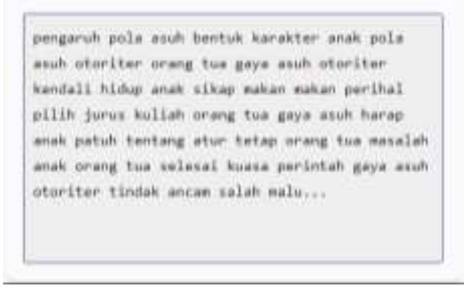
As shown in Figure 6, the omission of meaningless words was also conducted through a filtering process. The omitted words included the words dengan, seluruh, mulai, dari, yang, akan, nanti, dan seperti.



```
pengaruh pola asuh pembentukan karakter anak
pola asuh otoriter orang tua gaya pengasuhan
otoriter mengendalikan kehidupan anak bersikap
makanan dimakan perihal memilih jurusan kuliah
orang tua gaya pengasuhan mengharapkan anak
patuh menentang peraturan ditetapkan orang tua
permasalahan anak orang tua menyelesaikannya
kekuasaan perintah gaya pengasuhan otoriter
tindakan mengancam menyalahkan mempermalukan...
```

Figure 6 the omission of meaningless words (stopword)

Figure 7 shows the results of the stemming process, where the data changes to form basic words. Words that experience changes in examples included pembentukan to bentuk, pengasuhan to asuh, mengendalikan to kendali, kehidupan to hidup, and bersikap to sikap.



pengaruh pola asuh bentuk karakter anak pola asuh otoriter orang tua gaya asuh otoriter kendali hidup anak sikap makan makan perihal pilih jurus kuliah orang tua gaya asuh harap anak patuh tentang atur tetap orang tua masalah anak orang tua selesai kuasa perintah gaya asuh otoriter tindak ancam salah malu...

Figure 7 The results of the stemming process

4.2. Feature Extraction

Feature extraction in this research was carried out using the TFIDF method. Feature extraction aimed to obtain features from the data. Based on word frequency in the data as well as word frequency across the entire dataset, the TFIDF approach generated data features. The distribution of words across the data (common words) was represented using IDF. The frequency of words with less common word influence was represented by TFIDF.

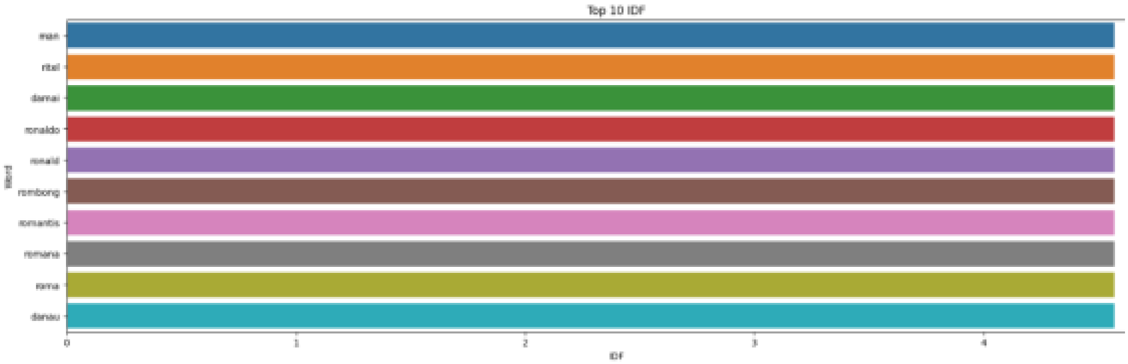


Figure 8 The highest IDF scores for ten features/word

The resulting features for each data were 2048 features/attributes. Ten words with the highest IDF scores can be seen in Figure 8. The highest TFIDF scores for ten features/words on each topic are shown in Figures 9 to 13.

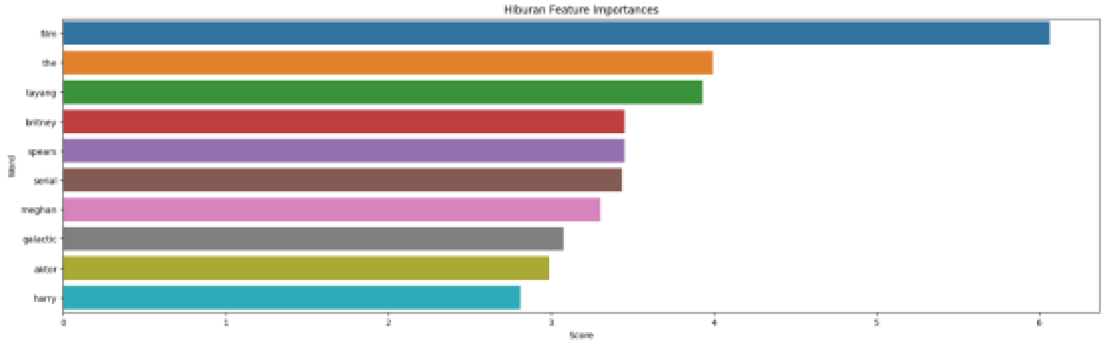


Figure 9 The ten highest weighted features for Entertainment Topic

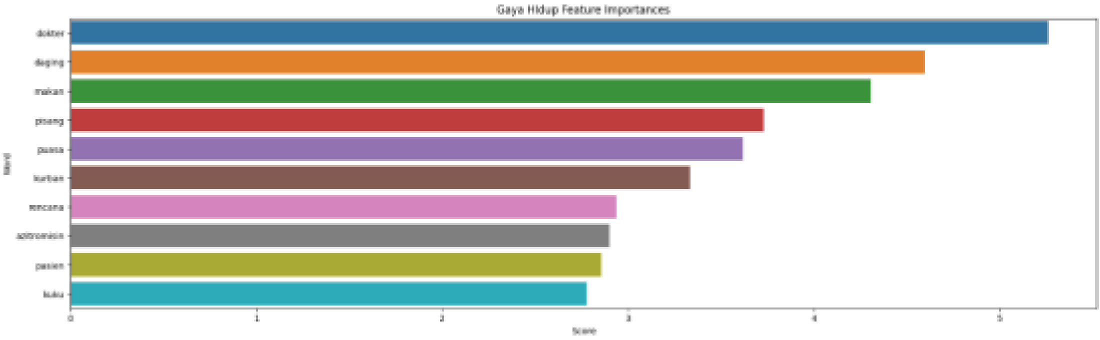


Figure 10 The ten highest weighted features for Lifestyle Topic

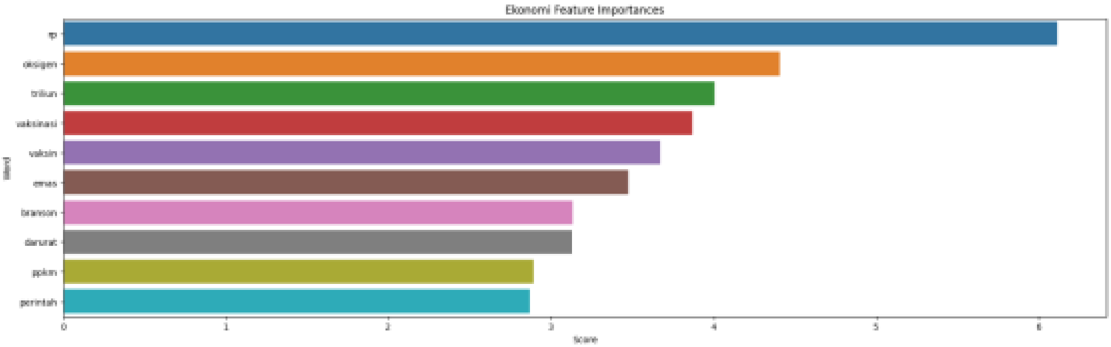


Figure 11 The ten highest weighted features for Economic Topic

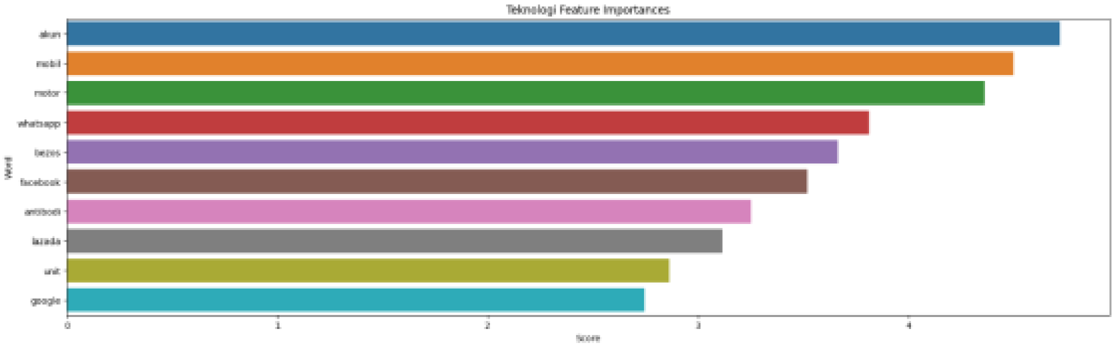


Figure 1 The ten highest weighted features for the Technology Topic

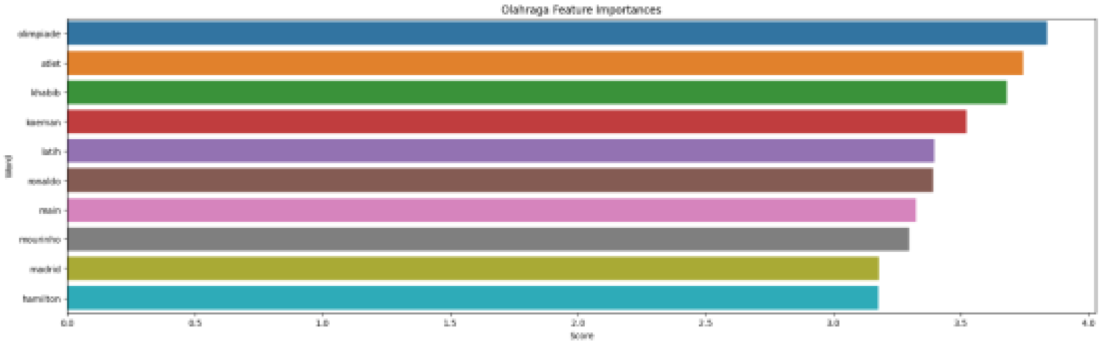


Figure 2 The ten highest weighted features for the Sport Topic

Figure 9 shows the ten highest weighted features for Entertainment Topic. Figure 10 shows the ten highest weighted features for Lifestyle Topic. Figure 11 shows the ten highest weighted features for Economic Topic. Figure 12 shows the ten highest weighted features for Technology Topic. Figure 13 shows the ten highest weighted features for the Sport Topic.

4.3. Feature Selection

Feature selection in this research aimed to select features based on their relevance to the topic. It was carried out using the ANOVA method, which utilized variations between the average feature/attribute data from various classes/groups. The ten features resulting from the feature selection process can be seen in Figure 14. The size of the data features after the feature selection process was 215.

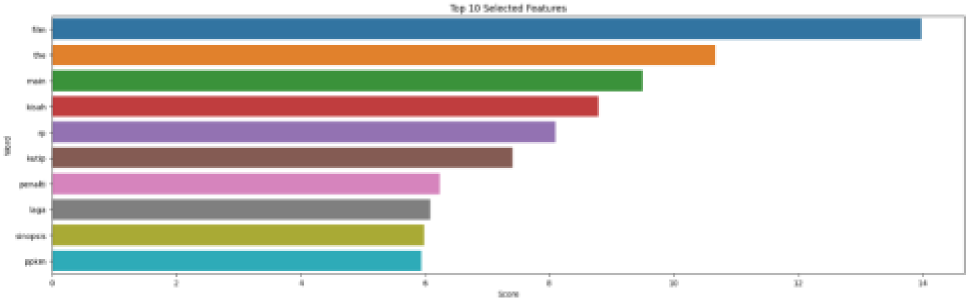


Figure 143 The ten highest weighted features from the feature selection process

4.4. Feature Reduction

Dimensional reduction aimed to reduce the data feature dimension, which used the PCA method in this research. PCA was obtained based on the principal data components with the highest value variation. The change in feature size resulted from a dimension reduction to 3 features. Visualization of the feature distribution resulting from feature selection is shown in Figure 15. Entertainment and Sports Topic features have separate feature sections from other features in one direction. Meanwhile, Economic Topic features have a section separated from other features in two directions. Lifestyle Topic features with a slightly separate section from other features, while Technology Topic features whose data were still combined with other features.

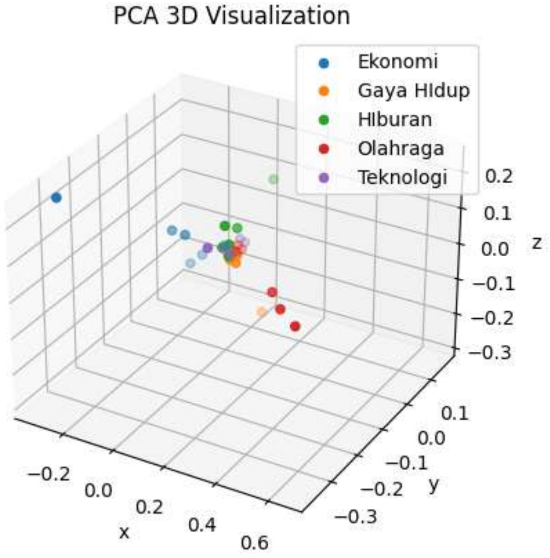


Figure 154 The feature distribution resulting from dimensional reduction

4.5. The Model Testing Result

The test was carried out on a 30% ratio data test or 30 article data. Based on multiple experiments with various parameter combinations, the optimum model was found. The best model for using the XGBoost method in this study shows the highest accuracy of 77%, with a precision of 81% and a recall of 77%. The confusion matrix of the test model obtained is shown in Figure 16. The misclassification for Economic and Sports Topics occurred for 1 article each, whereas 2 articles for Entertainment Topics and 3 articles on Technology Topics are misclassified. Economic Topics are incorrectly classified as Entertainment Topics, Sports Topics are incorrectly classified as Lifestyle Topics, and Entertainment Topics are incorrectly classified as Lifestyle Topics. Technology Topics that encountered misclassification of 2 articles were classified into Entertainment Topics and 1 article into Economic Topics.

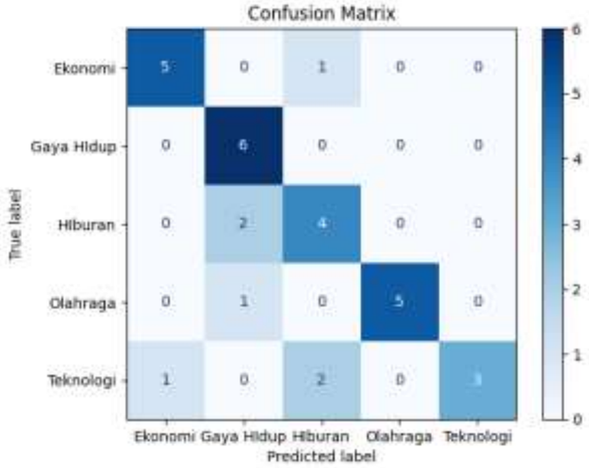


Figure 165 The Confusion matrix of test result

The ensuing model combination of parameters includes *max_depth*: 2; *min_child_weight*: 1; *gamma*: 0.0; *subsample*: 0.6; *colsample_bytree*: 0,1; *learning_level*: 0.2; and *n_estimator*: 100.

5. Conclusion

The conducted research aims to classify text data by employing the XGBoost classification method. Text mining was applied to classify text data through several stages, namely as follows. There was preprocessing, feature extraction using the TFIDF method, feature selection using the ANOVA method, feature dimension reduction using the PCA method, and classification using the XGBoost method. A text classification system was developed to classify text of articles. The highest accuracy obtained from the test is 77%, with a precision of 81% and a recall of 77%. Misclassification occurred on the Topic of Economics, Sports, and the most errors on the Topic of Technology.

References

[1] Lestari, N. M. A., & Sudarma, M. 2017. Perencanaan Search Engine E-commerce dengan Metode Latent Semantic Indexing Berbasis Multiplatform. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*. https://doi.org/10.24843/lkjiti.2017.v08.i01.p04

[2] Devi, A. S., Putra, I. K. G. D., & Sukarsa, I. M. 2015. Implementasi Metode Clustering DBSCAN pada Proses Pengambilan Keputusan. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*. https://doi.org/10.24843/lkjiti.2015.v06.i03.p05

[3] Zaki, M. J. & Meira W. 2014. DATA MINING Fundamental Concepts and Algorithms (2nd ed.) [Online]. Available: https://dataminingbook.info/

[4] Deng, X., Li, Y., Weng, J., & Zhang, J. (2019). Feature selection for text classification: A review. *Multimedia Tools and Applications*, *78*(3). https://doi.org/10.1007/s11042-018-6083-5

[5] Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. 2019. Comparing automated text classification methods. *International Journal of Research in Marketing*, *36*(1), 20–38. https://doi.org/10.1016/j.ijresmar.2018.09.009

[6] Liu, G., & Guo, J. 2019. Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing*, *337*. https://doi.org/10.1016/j.neucom.2019.01.078

[7] Stein, R. A., Jaques, P. A., & Valiati, J. F. 2019. An analysis of hierarchical text classification using word embeddings. *Information Sciences*, *471*. https://doi.org/10.1016/j.ins.2018.09.001

[8] Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. 2019. Text classification algorithms: A survey. *Information (Switzerland)*, *10*(4), 1–68. https://doi.org/10.3390/info10040150

[9] Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. https://doi.org/10.1145/2939672.2939785

[10] Mu, F., Gu, Y., Zhang, J., & Zhang, L. 2020. Milk source identification and milk quality estimation using an electronic nose and machine learning techniques. In *Sensors (Switzerland)* (Vol. 20, Issue 15, pp. 1–14). MDPI AG. https://doi.org/10.3390/s20154238

[11] Li, M., Fu, X., & Li, D. 2020. Diabetes Prediction Based on XGBoost Algorithm. *IOP Conference Series: Materials Science and Engineering*, *768*(7). https://doi.org/10.1088/1757-899X/768/7/072093

[12] Nyoman, N., Pinata, P., Sukarsa, M., Kadek, N., & Rusjayanthi, D. 2020. Prediksi Kecelakaan Lalu Lintas di Bali dengan XGBoost pada Python. *JURNAL ILMIAH MERPATI*, *8*(3), 188–196. https://ojs.unud.ac.id/index.php/merpati/article/download/63592/37798/

[13] Sukarsa, I. M., Pandika Pinata, N. N., Kadek Dwi Rusjayanthi, N., & Wisswani, N. W. 2021. Estimation of Gourami Supplies Using Gradient Boosting Decision Tree Method of XGBoost. *TEM Journal*, *10*(1). https://doi.org/10.18421/TEM101-17

[14] Can, R., Kocaman, S., & Gokceoglu, C. 2021. A comprehensive assessment of XGBoost algorithm for landslide susceptibility mapping in the upper basin of Ataturk dam, Turkey. *Applied Sciences (Switzerland)*, *11*(11). https://doi.org/10.3390/app11114993

[15] Shakhovska, K., Shakhovska, N., & Veselý, P. 2020. The sentiment analysis model of services providers' feedback. *Electronics (Switzerland)*, *9*(11), 1–15. https://doi.org/10.3390/electronics9111922

[16] Jun, S. 2021. Technology integration and analysis using boosting and ensemble. *Journal of Open Innovation: Technology, Market, and Complexity*, *7*(1), 1–15. https://doi.org/10.3390/joitmc7010027

[17] Das, S., Datta, S., Zubaidi, H. A., & Obaid, I. A. 2021. Applying interpretable machine learning to classify tree and utility pole related crash injury types. *IATSS Research*, *45*(3), 310–316. https://doi.org/10.1016/j.iatssr.2021.01.001

[18] Wang, P., Yan, Y., Si, Y., Zhu, G., Zhan, X., Wang, J., & Pan, R. 2020. Classification of Proactive Personality: Text Mining

[19] Thongsuwan, S., Jaiyen, S., Padcharoen, A., & Agarwal, P. 2021. ConvXGB: A new deep learning model for classification problems based on CNN and XGBoost. *Nuclear Engineering and Technology*, *53*(2). https://doi.org/10.1016/j.net.2020.04.008

[20] Jia, W., Sun, M., Lian, J. et al. 2022. Feature dimensionality reduction: a review. *Complex & Intelligent Systems*. 8, 2663–2693. https://doi.org/10.1007/s40747-021-00637-x