Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 277−285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2033

Multi-aspect Sentiment Analysis of Tiktok Application Usage Using FasText Feature Expansion and CNN Method

Rifki Alfian Abdi Malik*, Yuliant Sibaroni

Informatics Study Program, Telkom University, Bandung, Indonesia
Email: [1*]rifkialfian@student.telkomuniversity.ac.id, [2]yuliant@telkomuniversity.ac.id
Submitted: **02/08/2022**; Accepted: **23/08/2022**; Published: **30/08/2022**

Abstract-Among the many social media platforms that have emerged, TikTok is a platform that has the most significant number of subscribers compared to other platforms. However, not all reviews given by TikTok users are good reviews and reviews are often found with slang and not all reviews have real meaning, therefore sentiment analysis is needed for these problems. These reviews will later be analyzed for sentiment according to predetermined aspects, namely feature aspects, business aspects, and content aspects based on reviews written on the Google Play Store, using data crawling techniques and will pass the preprocessing and weighting stages. The weighting method used is Term Frequency-Inverse Document Frequency (TF-IDF). Then, the sentiment analysis process will use the Convolutional Neural Network (CNN) method, and feature expansion will be carried out to determine what words are interrelated with certain words. The purpose of this research is to analyze sentiment using Convolutional Neural Network and fastText feature expansion. The highest accuracy result is 87.74%.

Keywords: Multi Aspect, Sentiment Analysis, TikTok, Convolutional Neural Network, Google Play Store.

1. INTRODUCTION

In this age of rapidly developing technology, all kinds of things can be accessed very easily. Social media is one of the results of the development of internet-based technology, where users can continue to interact without the limitations of space and time that hindered human interaction in the past [1]. TikTok is one of the most popular social media platforms today. TikTok also includes a social networking application and music video platform where users can edit and share short video clips complete with filters and accompanied by music as support [2]. The TikTok app can be downloaded from the Google Play Store for Android users. Google Play Store provides a review function to rate apps, movies, and other services. With the review function, a lot of textual data is available. These text collections are a rich source of opinions, views, and emotions. A large number of reviews complicates the sentiment analysis process. Therefore, there is a need for a system that can perform sentiment analysis on aspects of these reviews and a feature expansion feature is needed to handle mismatches.

Sentiment analysis is the process of identifying sentiments that appear in a text by processing textual data to understand the opinions contained in a sentiment. Using sentiment analysis, the process of extracting information can be done to analyze the opinions contained in Google Play Store reviews written by users[3]. Sentiment analysis is a subset of computer text classification that intends to classify text based on its polarity, such as positive, negative, and neutral [4]. Therefore, analyzing the multi-faceted sentiment towards the TikTok app using Google Play Store reviews is very important.

Research [5] analyzed Naïve Bayes and Convolutional Neural Network methods using BOW extraction features and Word2Vex expansion features for text classification. It showed that the Convolutional Neural Network method with expansion features using Word2Vec got the highest accuracy at 91%. Furthermore, research [6] also analyzes multi-aspect sentiment on the Netflix application by combining SVM and LDA methods which get the highest f-1 score at 78.15%. The next research [7], analyzed the sentiment of text classification using fastText with a dataset of movie review sentences, and fastText got higher accuracy than word2vec by 85.2%.

Researcher [8] combined the K-Means Algorithm and Convolutional Neural Network in Android Application Classification based on permisson, and got the highest accuracy result of 92.23%. Research [9] compares deep learning methods with machine learning methods, the deep learning method used is Convolutional Neural Network while the machine learning methods used are SVM, KNN, and GSD. The accuracy results obtained in analyzing twitter sentiment with a general Indonesian dataset is 81.4%, this accuracy is the highest compared to other machine learning accuracies. Furthermore, in research [10] The final test accuracy results obtained are obtained accuracy in recognizing vegetable types of 98.1% with one of the best test results, namely the classification of corn vegetables with an accuracy of 99.98%.

Based on these studies, the motivation of this research is to conduct experiments to improve the accuracy value of previous studies by combining CNN and FastText. The dataset used in this research is TikTok application reviews obtained from the Google Play Store. In addition, the difference between the proposed and previous research is that it runs several scenario stages in its testing.

This research aims to build a sentiment analysis system on Google Play Store reviews of the TikTok application with the Convolutional Neural Network method and FastText Feature Expansion. And to find out the performance value of combining the CNN method and FastText feature expansion in calibrating the multi

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 277−285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2033

aspects of sentiment analysis of the TikTok application. The expected result is the performance of sentiment classification accuracy in every aspect consisting of Features, Business, and Content.

2. RESEARCH METHODS

2.1 Research Stages

The system overview shows the flow that runs on the system, as depicted in the following flowchart in Figure 1.

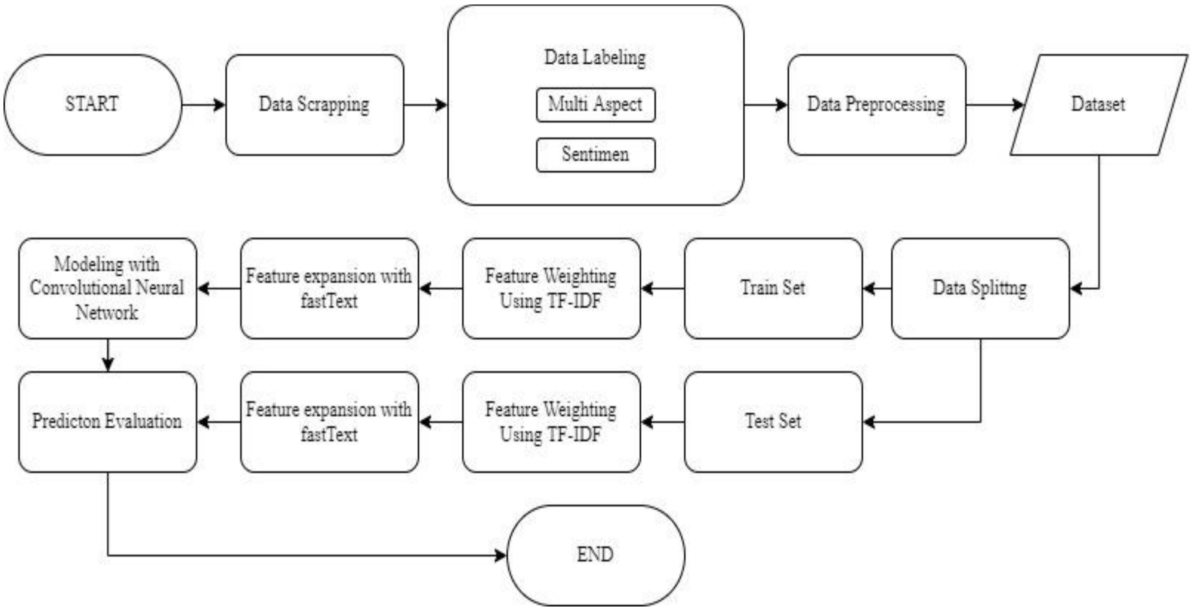

Figure 1. CNN Classification Model Process

Based on Figure 1, there is a flow of stages that will be carried out. The first stage is to collect data using google_play_scrapper to be used as a dataset for sentiment analysis models using the Convolutional Neural Network method. The second stage is to do multi-labeling based on Multi aspects and Sentiment. Then proceed with performing several pre-processing stages on the dataset that has been labeled. The next stage the dataset is split into train data and test data. Then weighting each word in the dataset using TF-IDF. Next, the feature expansion stage is carried out with fasText to determine what words are interrelated with certain words. Then train the model based on the training data with the Convolutional Neural Network method. The last stage is to evaluate the performance of the model that has been made.

2.2 Data Scrapping

Data scraping is the process of extracting data from a source that can be stored and processed. The data scraping process in this research is carried out using a library in python called google_play_scraper, which has a special function to get data directly from the Google Play Store website. The data used in this study are 10,000 TikTok application reviews obtained on March 23, 2022.

2.3. Data Labeling

The dataset is labeled manually. There are aspect labels and sentiment labels that serve as identifiers for which aspect and sentiment each review is categorized into. For aspect labels, there are three categories, namely there are feature aspects, business aspects, and content aspects. Likewise, for sentiment labels, there are three categories, namely positive, negative and neutral. The determination of table values in this study is shown in table 1.

Table 1. Determination of label value

Label	Description
-1	Negative value in an aspect
0	Not included in aspect
1	Positive value in an aspect

The labeling of each aspect is done based on the related vocabulary. For example, on "video nya bagus bagus" the data is included in the content aspect but not in the features and business aspects, and the determination of the vocabulary of each aspect in this study is shown in Table 2.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 277−285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2033

Table 2. Vocabulary Determination

Feature	Business	Content
Login	Uang	Video
Edit	Untung	Hiburan
Koin	Koin	Gambar
Follow	Poin	Informasi
Filter	Bonus	Media
Aplkasi	Saham	Viral
Profile	Investasi	Fyp (for your page)

The vocabulary used to label negative and positive sentiments in this study is shown in Table 3.

Table 3. Vocabulary Used

Positive	Negative
Baik	Jelek
Bagus	Kurang
Menarik	Susah
Semangat	Sampah
Menghibur	Rugi
Kenceng	Lemot
Mantap	Burik
Keren	Mengecewakan

2.3 Preprocessing

After labeling the data and before the data is processed for sentiment analysis. Data must be preprocessed to ensure the quality of the data is good before it is used during data analysis.

a. Data Cleaning
Data cleaning is the process of removing noise. Converts all letters to lowercase. This process is done to remove Hashtags, punctuation marks, usernames, numbers, and special characters can be seen in table 4.

Table 4. Data Cleaning Results

Input	Output
Saya suka banget aplikasi ini, tapi sekarang berubah dan jadi malas tiap posting video tidak pernah fyp dan follower selalu berkurang setiap harinya.	Saya suka banget aplikasi ini tapi sekarang berubah dan jadi malas setiap posting video tidak pernah fyp dan follower selalu berkurang setiap harinya

b. Case Folding
Case folding is a process to convert or remove all capital letters from textual data into lowercase letters (a-z). The result of case folding can be seen in table 5.

Table 5. Case Folding Results

Input	Output
Saya suka banget aplikasi ini tapi sekarang berubah dan jadi malas setiap posting video tidak pernah fyp dan follower selalu berkurang setiap harinya	saya suka banget aplikasi ini tapi sekarang berubah dan jadi malas setiap posting video tidak pernah fyp dan follower selalu berkurang setiap harinya

c. Tokenization
Tokenization is a process of fragmenting or cutting textual data into word parts called tokens. The result of tokenization can be seen in table 6.

Table 6. Tokenization Results

Input	Output
saya suka banget aplikasi ini tapi sekarang berubah dan jadi malas setiap posting video tidak pernah fyp dan follower selalu berkurang setiap harinya	‘saya’, ‘suka’, ‘banget’, ‘aplikasi’, ‘ini’, ‘tapi’ ‘sekarang’, ‘berubah’, ‘dan’, ‘jadi’, ‘malas’, ‘setiap’, ‘posting’, ‘video’, ‘tidak’, ‘pernah’, ‘fyp’, ‘dan’, ‘follower’, ‘selalu’, ‘berkurang’, ‘setiap’, ‘harinya’,

d. Stopwords Removal

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 277−285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2033

Stopwords removal is a process to remove non-topic words that are considered unimportant. This process helps in reducing less relevant features in the data. The stopwords used are built-in from the stopwords natural language toolkit, there are 757 stopwords that can be used, here is a table of 40 stopwords taken as samples.

Table 7. Stopwords Dictionary

Stopwords			
Ada	Kiranya	Membuat	Ungkapnya
Memihak	Makin	Buat	Supaya
Hingga	Segalanya	Diperlihatkan	Kembali
Bagaikan	Harusnya	Olehnya	Sendirinya
Demikian	Besar	Berlalu	Keinginan
Tentulah	Sambil	Ditunjuk	Begitu
Ingin	Seharusnya	Akankah	Kasus
Anda	Meyakini	Cukupkah	Kitalah
Berkali-kali	Beri	Kapanpun	Kepada
Antaranya	Dari	Tegasnya	Bagi

The results of the stopwords removal process can be seen in table 8.

Table 8. Stopwords Removal Results

Input	Output
‘saya’, ‘suka’, ‘banget’, ‘aplikasi’, ‘ini’, ‘tapi’ ‘sekarang’, ‘berubah’, ‘dan’, ‘jadi’, ‘malas’, ‘setiap’, ‘posting’, ‘video’, ‘tidak’, ‘pernah’, ‘fyp’, ‘dan’, ‘follower’, ‘selalu’, ‘berkurang’, ‘setiap’, ‘harinya’,	‘saya’, ‘suka’, ‘banget’, ‘aplikasi’, ‘sekarang’, ‘berubah’, ‘malas’, ‘setiap’, ‘posting’, ‘video’, ‘tidak’, ‘pernah’, ‘fyp’, ‘follower’, ‘selalu’, ‘berkurang’, ‘setiap’, ‘harinya’,

e. Stemming

Stemming is a process that transforms the words contained in the rootword by removing suffixes infixes, and confixes. The results of the stemming process can be seen in table 9.

Table 9. Stemming Results

Input	Output
‘saya’, ‘suka’, ‘banget’, ‘aplikasi’, ‘ini’, ‘tapi’ ‘sekarang’, ‘berubah’, ‘dan’, ‘jadi’, ‘malas’, ‘setiap’, ‘posting’, ‘video’, ‘tidak’, ‘pernah’, ‘fyp’, ‘dan’, ‘follower’, ‘selalu’, ‘berkurang’, ‘setiap’, ‘harinya’,	‘saya’, ‘suka’, ‘banget’, ‘aplikasi’, ‘sekarang’, ‘ubah’, ‘malas’, ‘setiap’, ‘posting’, ‘video’, ‘tidak’, ‘pernah’, ‘follower’, ‘selalu’, ‘kurang’, ‘setiap’, ‘harinya’,

2.4 TF-IDF Feature Extraction

Feature extraction is the process of converting or, in other words extracting documents in text format into features that can be easily processed by machine learning classification techniques. The feature extraction technique is one of the important techniques in data mining and text classification that calculates the value of features in documents [11]. In this research, TF-IDF is used for weighting and feature extraction. This weighting and feature extraction method is widely used, especially for information retrieval. In TD-IDF, term frequency (TF) is normalized with inverse document frequency (IDF). This normalization determines the weight of terms that frequently appear in a document. Basically, the document is converted into a weight that is calculated based on the number of occurrences in the document [12].

In [11], TF-IDF scored mostly higher than BM25 method in terms of f1-score. This is because the TF-IDF method uses two different tools, namely the local formula (TF) and the global formula (IDF), which give effective results [11]. This research uses TfidfVextorizer from Sckit-learn.org.TfidfVectorizer provides fit and transform features. The TF-IDF model learns vocabulary and document frequency for each word in the training data and then transforms it for train data and test data [13]. TF-IDF calculation in Scikit-learn is defined as follows [13]:

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 277−285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2033

$$tfidf(k,T) = tf(k,T)\ x\ idf(k) \tag{1}$$

$$idf(k) = log\frac{1+n}{1+df(k)} + 1 \tag{2}$$

Where tf (k,T) is the occurrence of word k in review T, n is the total review documents, and df(k) is the number of review documents containing word k. then the results of the TF-IDF vector are normalized with the Euclidean norm with the following calculation [13].

$$v_{norm} = \frac{v}{\sqrt{{v_1}^2+{v_2}^2+\cdots+{v_n}^2}} \tag{3}$$

2.5 Fast Text Feature Expansion

The next stage is feature expansion. Feature expansion is done using fast Text. First, fast Text is trained using the news dataset to generate a similarity corpus that can get the similarity of a word. For example, it can be seen in table 10 a list of words that have similarities with the word "saham" which has been sorted according to its similarity value.

Table 10. Example Similarity word of the word "saham"

Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Sahama	myrxpsaham	sahamnya	sahampada	sahan

Fast Text calculates the similarity score using cosine similarity [14]. The cosine similarity calculation for vectors A and B can be written as follows: fast Text calculates the similarity score using cosine similarity [14]. The cosine similarity calculation for vectors A and B can be written as follows:

$$Cosine\ Similarity(A,B) = \frac{A \cdot B}{||A||\ ||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{4}$$

Where the notation ||A|| is the definition of Euclidean normalization. After the model from fast Text training can get similar words, continue with the feature expansion process on the representation vector obtained from the previous feature extraction. Features whose vector value is 0 are replaced with the value of similar words that appear in Google Play Store reviews. The steps taken in this feature expansion process refer to research [15].

As an illustration, there is a review with the content "myxrpsaham sedang naik drastis" and there is a "saham" feature with a vector value of 0. The "saham" feature is expanded by first finding similarwrods as listed in table 10 . Furthermore, it is found that "saham" has similarities with "myxrpsaham" and the word "myxrpsaham" is contained in the review, so the vector value of "saham" is replaced with the vector value of "myxrpsaham".

2.6 Convolutional Neural Network

Deep learning is a part of Artificial Neural Network that was first introduced by Rina Dechter in 1986. Convolutional Neural Network (CNN) is a form of feed forward artificial neural network that is often used to model with input in the form of image or video data [16]. CNN is a type of Deep Neural Network due to its high network depth and is widely applied to image data. CNNs are inspired by biological processes where connectivity patterns between neurons resemble the organization of the visual cortex in animals [17]. The architecture of CNN is divided into 2 major parts, namely the Feature Extraction Layer and the Fully-Connected Layer. As shown in Figure 2 below:

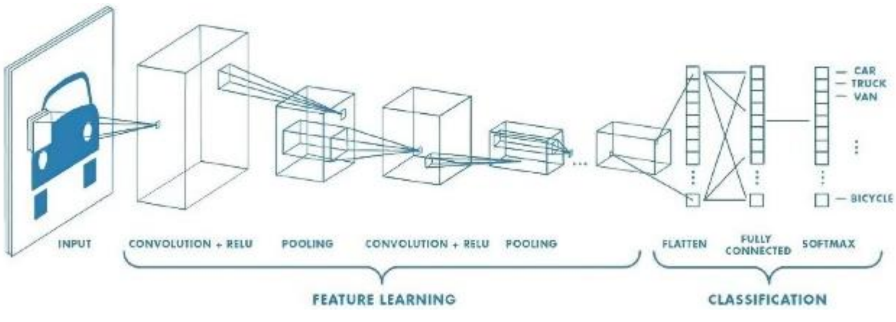

Figure 2. Architecture CNN

All the layers in the CNN are arranged in a stacked manner, like a sandwich consisting of bottom bread, vegetables, meat, cheese, tomato sauce, mayonnaise, chili sauce, and top bread. The input to the CNN has a 3-dimensional architecture: width, height, and depth. The width and height in the input image represent the

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 277−285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2033

dimensions of the image, while the depth represents the depth of the image, such as the Red, Green, Blue channels in an RGB image. The input will then enter the convolution layer, which is the core building block of CNN, where most of the computation is done in this layer. The results of the convolution process in the convolution layer will then be activated with ReLU in the activation layer which is useful for increasing the non-linearity of the decision function and the network as a whole without affecting the receptive field in the convolution layer. The Pooling Layer will then perform down sampling of the activated convolution results [18].

The value of the obtained from the Pooling Layer will then be converted into a vector called flatten which will then enter the FullyConnected Layer and then produce output from the results of the feed forward process on the CNN [19].

A convolution Layer is a layer that contains convolution operations. Most of the computation is in this layer [19]. This layer is also the main process underlying a CNN. Convolution is a mathematical term that means applying a function to the output of another function repeatedly. In image processing, convolution means applying a kernel to the image at all possible offsets [18]. The kernel moves from the top left corner to the bottom right corner as obtained by the convolution of the image shown in Figure 3 below:

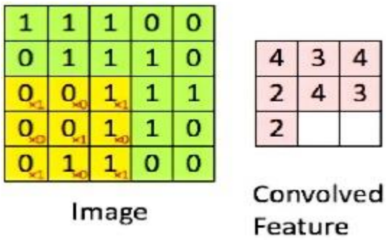

Figure 3. Convolution Process

The purpose of convolution is to extract the input image where the convolution process will produce a linear transformation of the input data according to the spatial information of the data. Formally, the convolution operation can be written with the following formula:

$$s(t) = (x * w)(t) \tag{5}$$

The function s(t) gives a single output in the form of a Feature Map, the first argument is the input which is x and the second argument was a filter while t as a pixel in the 2-dimensional image that will be replaced with i and j.

Max-Pooling is a down sampling process of reducing samples to maintain the size of the data from the previous convolution process. Max-Pooling on CNN for text is done by taking the largest value on the results of the convolution process that has been activated with ReLU before [20]. The following max-pooling equation with {c} is the value of the entire feature map value.

$$c = \max\{c\} \tag{6}$$

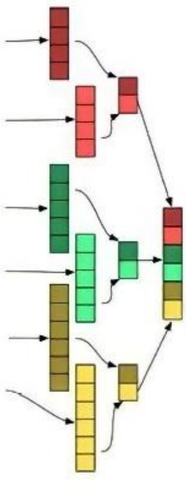

Figure 4. Max Pooling Layer

3. RESULTS AND DISCUSSION

This research conducted 2 scenarios, namely the preprocessing scenario and the feature expansion scenario. The dataset used at this stage is also explained.

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 277−285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2033

3.1 Dataset

The data for training and testing the model amounted to 10,000 Indonesian reviews. This dataset is divided into three labels that have been combined with multiple aspects of sentiment analysis. The distribution of dataset labels can be seen in table 11.

Table 11. Dataset Result

Data	Feature	Content	Business	Description
Aplikasinya bagus.	1	0	0	This is positive in terms of features, as there is pre-defined vocabulary in terms of content.
Bisa mendapatkan bonus koin jika sering upload, sama ngeditnya mudah banget.	1	0	1	There is a vocabulary in the feature called "edit" that has a positive score of "1", and there is a vocabulary called "bonus" that has a positive score of "1".
Kecewa, kenapa fitur live nya hilang padahal umur dan follower sudah memenuhi syarat Tolong di perbaiki lagi.	-1	0	0	Negatively included in the features aspect as there is pre-defined vocabulary in the business aspect.

For feature expansion with fastText, the fastText corpus was built with a news dataset obtained from research [15] which contains Indonesian news taken from mainstream media. The news corpus used amounted to 97794 articles, and the media that became the source of the dataset can be seen in the table 12.

Table 12. Article Source Data Distribution

Source	Total number of articles
Cnnindonesia.com	24308
Detik.com	1750
Kompas.com	7400
Republika.com	43536
Sindonews.com	11110
Tempo.co	9690
Total	97794

3.2 Ektraksi Fitur TF-IDF

From 10,000 documents, 7643 features were obtained after the TF-IDF feature extraction process was carried out, here is a partial table of the results of the TF-IDF Extraction Features carried out:

Table 13. Aspect Feature Result

Top Feature		
-1	0	1
Jelek	Apk	Apk
Live	Aplikasi	Aplikasi
Fitur	Bagus	Bagus
Update	Banget	Banget

Table 14. Aspect Business Result

Top Feature		
-1	0	1
Bonus	Bayar	Cuan
Dana	Bonus	Duit
Event	Dana	Hasil
Kode	Dapet	Uang

Table 15. Aspect Content Result

Top Feature

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 277−285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2033

-1	0	1
Fyp	Fyp	Gabut
Live	Hibur	Hibur
Video	Konten	Manfaat
Konten	Manfaat	Senang

The result of this TF-IDF feature extraction process is used for feature expansion.

3.3 Scenario 1 Testing Preprocessing

Three tests were carried out, namely testing with full preprocessing, then testing without stemming, and the last one is without stopwords. This test is conducted to identify whether there is a significant difference in the results of each process in each data preprocessing. The results of this test can be seen table 16 :

Table 16. Scenario Preprocessing

Scenario	Accuracy (%)			Average (%)
	Feature	Content	Business	
Full Preprocessing	81.22	93.82	88.18	87.74
No Stop words	80.24	92.53	88.64	87.13
No Stemming	81.30	93.21	88.12	87.54

Table 13 shows that the full preprocessing scenario gets an average accuracy of 87.74%. Next, we will test feature expansion using Kfold k = 5 and full preprocessing dataset.

3.4 Scenario 2 Testing Fast Text Feature Expansion

The 2nd scenario test was conducted to compare CNN models that have implemented feature expansion with CNN models without feature expansion implementation. Feature expansion is performed on top 1, 5, and 10 features with similarity corpus built with a news dataset. Top 1 feature means taking one word that has the highest similarity value to a word from the fastText corpus and so on for each top-n feature.

Table 17. Feature Expansion with Fast Text Scenario

Scenario	Accuracy (%)			Average (%)
	Feature	Content	Business	
Baseline	81.22	93.82	88.18	87.74
Top-1 Feature	81.36 (+0.14)	92.54 (-1.28)	88.8 (+0.62)	87.56
Top-5 Feature	80.31 (-0.91)	92.98 (-0.84)	88.34 (+0.16)	87.21
Top-10 Feature	80.13 (-1.09)	93.14 (-0.68)	87.96 (-0.22)	87.07

Based on the results of scenario 2 testings that have been done. The highest accuracy obtained in the feature aspect is obtained by using the top-1 feature scenario, which is 81.36%. Using the top-1 feature can improve the performance of the CNN model created by 0.14% compared to the baseline feature aspect. While testing scenario 2 on the content aspect does not improve the accuracy obtained, the highest accuracy obtained is 93.14% using the top-10 feature. These results are lower than the baseline results with a difference of 0.68%. For the business aspect, the highest accuracy is obtained using the top-1 feature, which is 88.8%. Using the top-1 feature can improve the performance of the CNN model created by 0.82% compared to the baseline business aspect.

4. CONCLUSION

This research has conducted feature expansion with fast Text in multi-aspect sentiment analysis using 2 scenarios. The two scenarios include preprocessing testing and feature expansion testing. In the first test to determine the baseline, preprocessing testing was tested without stemming and without stop words. In the second test, the feature expansion scenario was carried out with the best dataset from the previous scenario added with model evaluation using 5-fold cross-validation. The feature expansions used include top-1 features, top-5 features, and top-10 features. The results of testing scenario 1 obtained the highest average accuracy of the three aspects of 87.74%, which was obtained by testing full preprocessing. The preprocessing stage affects the performance of the CNN model created. It is proven that the full preprocessing testing stage provides better results compared to preprocessing testing without stemming and preprocessing testing without stop words. Meanwhile, the results of scenario 2 testing obtained the best results with the highest average accuracy of 87.56% using the fast Text expansion feature from the top-1 feature. The use of feature expansion with fast Text in the CNN model affects the resulting performance. By using top-1 features, the accuracy obtained in the feature aspect and business aspect has increased compared to the baseline, but in the content aspect, the accuracy

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 277−285
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2033

has decreased by 1.28%, whereas the top-1 feature accuracy is 92.54%, while the baseline accuracy is 93.82%. Suggestions for further research can be made by experimenting with other neural network methods with different feature expansions in analyzing multi-aspect sentiment

REFERENCES

[1] F. Atefeh and W. Khreich, "A survey of techniques for event detection in Twitter," *Comput. Intell.*, vol. 31, no. 1, pp. 133–164, 2015, doi: 10.1111/coin.12017.
[2] W. N. Aji, "Aplikasi Tiktok Sebagai Media Pembelajaran Bahasa dan Sastra Indonesia," *Pros. Semin. Nas. Pertem. Ilm. Bhs. dan Sastra Indones.*, vol. 431, pp. 431–440, 2018.
[3] I. M. B. S. Darma, R. S. Perdana, and Indriati, "Penerapan Sentimen Analisis Acara Televisi Pada Twitter Menggunakan Support Vector Machine dan Algoritma Genetika sebagai Metode Seleksi Fitur," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 3, pp. 998–1007, 2018, [Online]. Available: http://j-ptiik.ub.ac.id.
[4] I. F. Ramadhy and Y. Sibaroni, "Analisis Trending Topik Twitter dengan Fitur Ekspansi FastText Menggunakan Metode Logistic Regression," *J. Ris. Komputer)*, vol. 9, no. 1, pp. 2407–389, 2022, doi: 10.30865/jurikom.v9i1.3791.
[5] J. Panthati, J. Bhaskar, T. K. Ranga, and M. R. Challa, "Sentiment Analysis of Product Reviews using Deep Learning," *2018 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2018*, pp. 2408–2414, 2018, doi: 10.1109/ICACCI.2018.8554551.
[6] A. R. Abelard and Y. Sibaroni, "Multi-aspect sentiment analysis on netflix application using latent dirichlet allocation and support vector machine methods," *J. Infotel*, vol. 13, no. 3, pp. 128–133, 2021, doi: 10.20895/infotel.v13i3.670.
[7] N. Nedjah, I. Santos, and L. de Macedo Mourelle, "Sentiment analysis using convolutional neural network via word embeddings," *Evol. Intell.*, pp. 2–6, 2019, doi: 10.1007/s12065-019-00227-4.
[8] T. Novriansyah Turnip, P. O. Manik, J. H. Tampubolon, P. Adi, and P. Siahaan, "Klasifikasi Aplikasi Android Menggunakan Algoritme K-Means Dan Convolutional Neural Network Berdasarkan Permission Android App Classification Using K-Means and Convolutional Neural Network Algorithms Based on Permission," vol. 7, no. 2, pp. 399–406, 2020, doi: 10.25126/jtiik.202072641.
[9] Sartini, *Analisis Sentimen Twitter Bahasa Indonesia Menggunakan Algoritma Convolutional Neural Network*. 2020.
[10] M. J. Akbar, M. W. Sardjono, M. Cahyanti, and E. R. Swedia, "Perancangan Aplikasi Mobile Untuk Klasifikasi Sayuran Menggunakan Deep Learning Convolutional Neural Network," *Sebatik*, vol. 24, no. 2, pp. 300–306, 2020, doi: 10.46984/sebatik.v24i2.1134.
[11] A. Kadhim, *Term Weighting for Feature Extraction on Twitter: A Comparison Between BM25 and TF-IDF*. 2019.
[12] A. I. Kadhim, "Survey on supervised machine learning techniques for automatic text classification," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 273–292, 2019, doi: 10.1007/s10462-018-09677-1.
[13] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[14] 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, doi: 10.1162/tacl_a_00051.
[15] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Feature expansion using word embedding for tweet topic classification," *Proceeding 2016 10th Int. Conf. Telecommun. Syst. Serv. Appl. TSSA 2016 Spec. Issue Radar Technol.*, no. October, 2017, doi: 10.1109/TSSA.2016.7871085.
[16] N. I. Widiastuti, "Deep Learning - Now and Next in Text Mining and Natural Language Processing," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 407, no. 1, 2018, doi: 10.1088/1757-899X/407/1/012114.
[17] S. B and U. A, *Data Mining dan Big Data Analytics: Teori dan Implementasi Menggunakan Python & Apache Spark.* Yogyakarta: Penebar Media Pustaka, 2018.
[18] F. H. Ihromi, "Ekstraksi Informasi Dokumen Karya Tulis Ilmiah Menggunakan Algoritma Convolutional Neural Network," *Unikom*, 2019, [Online]. Available: https://elibrary.unikom.ac.id/id/eprint/1511/.
[19] Suyanto, *Machine Learning: Tingkat Dasar dan Lanjut*. Informatika Bandung, 2018.
[20] Y. Zhang and B. Wallace, "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification," pp. 253–263, 2015, [Online]. Available: http://arxiv.org/abs/1510.03820.

