JURNAL INFORMATIKA
PROXIES

Vol.9, No.1, July, 2025

Comparison Of Cbow And Skip Gram On Sentiment Analysis Of Bus Agents Reviews Using CNN

Fidelis Raymond Jose Airudin[1], R. Setiawan Aji Nugroho S.T., MCompIT., Ph.D.[2]

[1] Teknik Informatika Fakultas Ilmu Komputer, Universitas Katolik Soegijapranata
[2] Teknik Informatika Fakultas Ilmu Komputer, Universitas Katolik Soegijapranata, nugroho@unika.ac.id

Corresponding Author Email: nugroho@unika.ac.id

https://doi.org/10.24167/proxies.v9i1.13281

Received: 2025-02-20
Revised:2025-10-22
Accepted: 2025-10-22
Available online:2025-10-23

Keywords:
Intercity Bus, Sentiment Analysis, Word2Vec, CNN, CBOW, Skip Gram)

ABSTRACT

Inter-city buses are one of the most popular modes of transportation in Indonesia. Thanks to the massive development and construction of highways by the Indonesian government, today almost every city on the island of Java is connected by toll roads. This makes traveling from one city to another faster, more comfortable, and free from obstacles. This opens up great opportunities for inter-city bus entrepreneurs to attract more passengers. Currently, inter-city bus companies are aggressively rejuvenating their fleet and adding attractive facilities to passengers, so that prospective bus passengers currently have many bus options to choose from. This of course creates a problem especially for new bus users because there are too many options to choose from. If prospective bus users are wrong in choosing a bus, besides the material can be a victim, of course the experience of bus users will also be a victim. Therefore, it is necessary to create a system to detect what is the most dominant sentiment expressed by bus users who have ridden the bus before through a review. To create a system to analyze sentiment, researchers use the Convolutional Neural Network algorithm where for the input layer researchers use a matrix representation of words formed by the word embedding method using the Word2Vec algorithm. As a final result, researchers display the results of the Precision, Recall, and F1-Score calculations which show the quality of the sentiment analysis model created.

1. INTRODUCTION

Intercity bus is one of the modes of public transportation that is often used by the public to travel from one city to another besides trains and planes. Almost every city in Indonesia, especially the island of Java, can be reached by inter-city buses. Thanks to the massive toll road infrastructure development carried out by the Indonesian government, the travel time required every time you travel outside the city is significantly reduced. According to Direktorat Jenderal Perhubungan Darat Kementerian Perhubungan (Kemenhub) data [1], there are currently 346 registered inter-city bus companies in Indonesia. The data of 346 intercity bus companies registered in Kemenhub by 2020 can illustrate that such numbers are not small. With so many inter-city bus options that can be chosen by the public, of course, it has the potential to cause confusion when choosing which inter-city bus to ride, especially bus passengers who are using inter-city buses for the first time.

Currently, many bus operators register their representative ticket sales agents on Google Maps, which can be used for marketing because on Google Maps, we can provide information such as addresses, opening hours, contact or telephone numbers, website URLs, and other information that can make it easier for passengers to find out more about the bus operator. On Google Maps there is a feature where we can write a review about a place we have visited, and we can write our opinion on that place based on the experience we got from that place. Of course, this feature is also available at the location of bus ticket sales agents listed on Google Maps. From what researchers got after making observations at several locations of bus ticket sales agents registered on Google Maps. This review feature is often used by bus passengers to write their opinions about the services of bus ticket sales agents and also the crew of the bus they are traveling on based on their experiences. Therefore, we can use this feature to find out whether the credibility of intercity bus operators is good or bad based on the sentiments contained in the reviews given by reviewers.

To find out the sentiment in a sentence from the review data so that we can identify the experience or opinion of a reviewer, in this study, researchers use a sentiment analysis method which is currently quite popular in processing or tasks related to analysis with data in the form of natural human language. Sentiment analysis is a method used to find out opinions and sentiments expressed by someone through words or from writing that is subjective in nature based on someone's psychology, experience, or perspective on something [2]. From previous studies, sentiment analysis can be carried out using a classification algorithm and can give good results. In this research the researcher chose a Convolutional Neural Network algorithm to perform sentence classification to find out whether the sentence or review given to the bus company has a positive, neutral, or negative connotation.

In general, convolutional neural networks are commonly used to classify images using the kernel and convolution technique to extract and preserve regional features of images which is then used as the input layer in a neural network. Because in this study the authors used text data, the text needed to be converted into a matrix representation so that it could be processed by CNN. By utilizing word embedding technology where sentences or text will be converted into matrix representation form, the researcher first converts sentences into a matrix representation form using word embedding technology which will then become the input layer of the convolutional neural network. Then to evaluate the CNN model that was built, the researcher uses an evaluation matrix to calculate precision, recall, and f1-score.

GOALS OF THE PAPER

The purpose of this research is to implement the Convolutional Neural Network algorithm to classify sentiment from bus reviews to identify the credibility owned by each of these intercity bus companies. Moreover, In this research, the researcher wants to discuss the two architectures contained in the Word2Vec algorithm, namely Continuous Bag of Words and Skip-gram, where these two architectures will later be used to convert sentences into vector representations which are then used as input to the input layer of the CNN algorithm.

LITERATURE STUDY

Word embedding is a technique that is quite popular in natural language processing which is used to represent words in the form of numeric vectors. Understanding natural language text data is one of the main tasks in the field of Natural Language Processing and is an essential part of the development of Artificial Intelligence, especially to make computers able to understand human natural language. As we know that the computer treats words in a sentence only as data without knowing what it means and the relationship between a word and another word. Therefore, so that the computer can find out the deeper meaning of these words, we need to change words that were previously in the form of plain text into vector form where the value will represent a certain meaning or aspect.

One of the methods to convert text data into a vector representation is to use one-hot encoding or commonly called one-hot vector. One-hot encoding will collect all the words contained in each document in the entire corpus and the frequency of occurrence of these words will be calculated in certain documents. For example, in the sentence "saya membeli tiket bus di agen bus rosalia indah" the vector representation would be {saya: 1, membeli: 1, tiket: 1, bus: 2, di: 1, agen: 1, rosalia: 1, indah: 1}. The vector may look simple because there is only one document, but if there are many documents in a corpus, the words contained in the vector will also increase which will burden the computer in processing text data. Therefore we need to use another alternative that is more efficient, namely by creating a word embedding model or what is commonly called a language model.

The goal of language modeling is to study the combined probabilities of a given word sequence occurring in a sentence [3]. The greater the probability value, it indicates that the word order is the word order that is often used. One method that is often used to study relations or relationships between words is n-gram, where n is a hyperparameter that is selected manually to determine how many words to calculate the probability of the relationship. If we choose n number equals to 2 then we will learn every 2 words contained in the document.

In the field of language modeling, there is one model that is quite popular for converting a word into a vector representation, namely Word2Vec. Word2Vec will produce a vector representation of a word based on the distribution of words contained in a corpus or document, which is the input of the model. The working principle of the Word2Vec algorithm is that a word that often appears simultaneously with other words in a sentence is considered to tend to have a similar or interrelated meaning.

Word2Vec is based on 2 architectures, namely Continuous Bag-of-Words (CBOW) and Skip-gram. These two architectures have different ways of working, in CBOW the target word will be predicted based on the context words around it, for example the sentence "saya suka pergi naik bus" then CBOW will try to predict the word "pergi" based on the context "saya suka naik bus" or the word "saya" based on the context "suka pergi naik bus" as well as for other words in the sentence. It is different from Skip-gram where what is predicted is the word context based on the target word. Skip-gram is useful if we want to study the relationship between words in a sentence. For example, in the sentence "saya suka pergi naik bus" if the word "pergi" is the target word then this word will be used as input to predict the surrounding context words, namely the world "saya", "suka", "naik", "bus" will be predicted.

In 2013, two architectures of word embedding models named Continuous Bag of Words and Skip-gram were introduced and proposed by Mikolov et al [4]. The two architectures were formed with the aim that we can calculate a continuous vector representation of a collection of words from very large data sets. The architectures or techniques that have been formed and introduced before the CBOW and Skip-gram are believed to have not been successfully trained using more than a few hundred million words with a fairly simple dimension of the word vector, which is between 50 to 100 dimensions. That is the one of the reasons for the development of these two architectures so that in the future there is a model that can be used to train high-quality word vectors from large data sets with billions of words, and with millions of words in the vocabulary.

Previously, Mikolov et al. observed and found that in the context of inflectional languages [4], that a noun can have several word endings which if traced for similar words in the subspace of an original vector space, allows us to find words that have similar endings. This is one of the basic ideas that the proposed technique of the two architectures will be trained and the quality of the model formed is seen from the quality of the resulting vector representation.

Sentiment analysis is a method that has an important role in various fields, especially used in developing decision-making systems [5]. Sentiment analysis is a computational study technique that studies opinions, sentiments, emotions, appraisals, and attitudes expressed by people towards an entity such as products, services, organizations, individuals, issues, events, and topics being discussed [2]. With the growth of technology that is growing exponentially as indicated by the emergence of new technology platforms such as social media, blogs, website-based discussion forums that can be accessed via the internet, allowing people from various parts of the world to communicate with each other and exchange information more easily, and activities The data will always be recorded every second so that the data stored is quite large or even very large. Through this phenomenon, since early 2000, sentiment analysis has developed into one of the most active research areas in Natural Language Processing (NLP) [6].

With the faster process of exchanging information that occurs through the internet network, it is easier for us to find out whether a product has good quality or not through feedback given by someone who has tried the product. Information regarding user feedback is usually provided in the form of reviews based on the reviewer's experience after using the product. Oktaviani et al [7], in their research, stated that reviews given on a product can help someone, especially a business owner, in evaluating improvements to the product they have. As research material, the author uses reviews provided by users of the Traveloka ticket and hotel sales application, which is quite popular, especially in Indonesia. To analyze the review data, the writer uses the sentiment analysis method by classifying each review into positive or negative sentiment groups using the Naïve Bayes Classifier algorithm and the text association method. To evaluate the model, the author uses a confusion matrix to calculate overall accuracy and Kappa Accuracy. The result is a comparison of 80% training data and 20% testing data, the authors get the best value for overall accuracy of 91.20% and the best value for Kappa Accuracy of 59.56%.

An online review of a product or service has an important role for buyers. The impression of someone's experience written in the form of a review is one of the factors that can influence someone to decide whether to buy the product or not. One example of an application that implements this online review feature is Traveloka, which is quite popular for searching and making hotel reservations. Dividing hotel customer preferences and satisfaction into several groups is an important step in analyzing customer behavior to improve the quality of hotel products and services [8]. The authors analyze some of the features needed by hotel customers using the reviews contained in the Traveloka application with the aim of finding important factors that can influence customers in choosing a hotel. To group desires or levels of satisfaction in order to identify customer segments, the author uses the HBDScan algorithm and the Machine Learning approach. HBDScan is a clustering algorithm in machine learning which is commonly used in segmenting markets. The author took as many as 1.8 million reviews from 3081 hotels on Traveloka in the Special Region of Yogyakarta. A total of 10000 data were randomly selected to be used as a dataset. As a result, a total of 44 clusters were formed where as many as 62.3% of customers gave positive reviews, 28.5% were negative reviews, and the remaining 9.3% were neutral reviews. Of the 44 clusters, 5 clusters were selected which had the bigger result among the others, and were then analyzed to determine what policies the hotel manager needed to implement. Each of the 5 big clusters contains reviews that assess the location, room facilities, general reviews, service, and food served.

Sentiment analysis if we use it well and formed with a well-trained model, there are many benefits that we can get with sentiment analysis. One of the reasons why researchers make this sentiment analysis model is so that in the future the model can be used by others to find references to which buses have good credibility to use when traveling. Systems that can provide references or choices to users are generally called recommendation systems. Isinkaye et al. [9] states that a recommendation system is an information filtering system that performs information overload filtering by filtering information that is considered vital from a large number of dynamically generated information based on preferences, user interests, or observations of user behavior towards an item. Collaborative Filtering is one of the algorithms commonly used for recommendation systems. The result of the algorithm is item recommendations by identifying other users who have the same interests or tastes. Jian, Shen et al. [10] stated that Collaborative Filtering is a technology that aims to study user preferences and make recommendations based on the historical behavior of users - users and community data or data from groups of users who have the same preferences and interests. To find out whether a user has the same interests as other users, we need to calculate the similarity between users based on the items that users give feedback, usually in the form of ratings or reviews.

The feedback given by the user to an item in the form of a rating or review turns out to be quite potential to cause a problem. Jiang et al. [11] states that currently there are many fake reviews or ratings given by users to an item on an e-commerce website. The fake reviews are divided into several categories such as, When the e-commerce holds a sale activity for e-commerce goods, it promises users to get some money if they give a high rating to certain items. Then another example is an e-commerce service provider hiring someone to provide an assessment of an item with a specific purpose. This of course will lead to an information bias that will harm other customers quite significantly. On the Amazon website there is a vote for each product review where the user will judge whether the review is helpful or not. Then Jiang et al. use the results of the vote as a parameter that the review is true and accurate in accordance with the experience felt by the reviewer.

RESEARCH METHODOLOGY

Dataset Collection

Researchers find and collect the dataset from 5 bus companies ticketing agents that have been registered on Google Maps. Researchers choose 5 samples of bus companies for the comparison. In the data collecting process, researchers use the web scraping method where researchers select the name of the bus company, its review, and the rating given by the bus user who gave the review.

In building and evaluating the sentiment analysis model, in this research, the researchers divided the training and testing data in a ratio of 20:80, 50:50, and 80:20. For each comparison ratio of training data and testing data, accuracy, precision, recall and f1 score will also be calculated.

Dataset Labeling

Because the data taken by researchers do not have labels, researchers need to label each data taken manually. The researcher was assisted by the researcher's mother and father to carry out the labeling and for comparison, the researcher also did it independently so that there was no bias in the information from the data the researcher took. Researchers divided them into three classes, namely positive, negative and neutral.

Research Methodology

Dataset Collection

Researchers find and collect the dataset from 5 bus companies ticketing agents that have been registered on Google Maps. Researchers choose 5 samples of bus companies for the comparison. In the data collecting process, researchers use the web scraping method where researchers select the name of the bus company, its review, and the rating given by the bus user who gave the review.
In building and evaluating the sentiment analysis model, in this research, the researchers divided the training and testing data in a ratio of 20:80, 50:50, and 80:20. For each comparison ratio of training data and testing data, accuracy, precision, recall and f1 score will also be calculated.

Dataset Labeling

Because the data taken by researchers do not have labels, researchers need to label each data taken manually. The researcher was assisted by the researcher's mother and father to carry out the labeling and for comparison, the researcher also did it independently so that there was no bias in the information from the data the researcher took. Researchers divided them into three classes, namely positive, negative and neutral.

No.	Sentiment	Explanation	Sentence
1.	Positive	Sentences that contain positive opinions or minimal criticism.	"Petugas cek in ramah, ruang tunggu luas walau tanpa AC tetapi ada kipas, untuk bus lumayan, ketepatan waktu oke, fasilitas toilet bersih dan wangi"
2.	Negative	Sentences that contain predominantly negative opinions or criticism.	"Bus nya sudah bagus, Cuma pelayanan pramugaranya kurang bagus, attitudenya kurang. Mohon ada perbaikan untuk dilatih lebih baik lagi"
3.	Neutral	Sentences that do not show strong opinions, or positive opinions balanced with minimal criticism, sentences that contain factual information without any clear opinions.	"Sekedar saran alangkah baiknya disediakan nomor rekening perusahaan atau mesin EDC karena banyak orang yang tidak bawa uang tunai"

Error! Reference source not found. is the reference used to do the labeling that has been agreed upon by the researcher and the two other annotators. Sentences that are categorized into positive sentiments are sentences that contain positive opinions or minimal criticism, meaning that if a review is dominated by sentences of praise, sentences that express satisfaction, or if a sentence has criticism but overall the reviewer shows satisfaction then the review will be included in the positive category. Furthermore, if a review contains sentences that express disappointment, dissatisfaction, or a criticism then the review will be put into the negative category. The neutral category is a category for sentences that contain factual information such as sentences that explain the location of the agent or the conditions around the ticket sales agent that do not show a dominant opinion between positive and negative.

Data Pre-Processing

Before entering the analysis process, one of the steps that needs to be done is to prepare the raw data that has previously been obtained and then convert it into a cleaner format so that the data has the right consistency and has good quality so that the model can learn effectively and the predictions produced are accurate and the model has good performance. The first step in pre-processing is tokenization. Tokenization is a process that will break each sentence into a list of words. This list of words will then be referred to as tokens and from this collection of tokens will later be used to create a one-hot encoding vector. Once the token set has been obtained, the next step is to clean special characters such as punctuation from the token set. These characters need to be removed because they do not mean anything to the model. After the punctuation and special characters are cleaned, what remains is a collection of words. These collections of words generally contain letters in uppercase format, therefore researchers need to change each letter to lowercase. The purpose of this process is to make the model read and interpret each word more easily and consistently. So that if there is a duplication of words that are distinguished because of the difference in uppercase and lowercase letters, the model will not distinguish them. After all the letters have been changed to lowercase, the next step is the process of removing common words that have no meaning, which are usually called stop words. After all stop words have been successfully removed, the next step is the process of changing the word into its root form, this process is called stemming. The purpose of the stemming process is to reduce words that have the same meaning so that the word can be analyzed as one entity.

Word Embedding

Word embedding is a representation of words in the form of vectors that have certain dimensions where the numbers contained in the vector can describe the semantic meaning and relationship between words in a sentence or document. Generally, the dimensions of the vector formed range from 100 to 300 dimensions, and these dimensions are fixed. The larger the dimension of the vector, the more detailed the representation of each word and the richer the semantic meaning, especially if the data used has many unknown or rare words. Word embedding is formed by learning the relationship between words in a large document. It is important to note that in order to produce a good vector representation, the data needs to be very large and contain a variety of words.

Word2vec is one of the commonly used word embedding algorithms. Word2Vec, is built on a neural network-based architecture. There are three main layers that make up the core of the Word2Vec model, namely the input layer where in this layer all words contained in the corpus are converted into a one-hot vector with the numbers 0 and 1 where the number 1 indicates a word that will be used as input from the Word2Vec model. Then the second layer is the hidden layer, Word2Vec has one hidden layer where its dimensions will store the representation of the input word into a vector. The size of this hidden layer is what we need to determine well based on the needs, goals, and tasks that we want to complete with the word embedding results later. The last is the output layer, just like the input layer, this output layer is a one-hot vector where the word to be predicted will be given the number 1 and the other 0. Because the output will be produced in the form of probability, the training process uses a softmax activation function.

Word2Vec has two architectures in it, each architecture has different characteristics and its use is also distinguished based on the task or goal to be achieved. The two architectures are Continuous Bag of Words and Skip Gram

Continuous Bag of Words

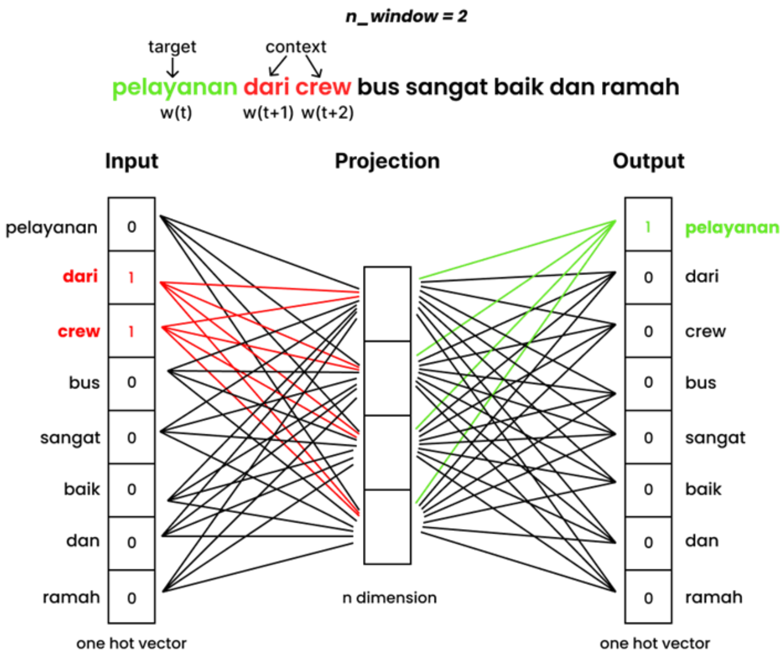

Figure 1. The Architecutre of Continuous Bag of Words

As we can see in the figure 0, the Continuous Bag of Words architecture, the Word2Vec model will try to predict the target word based on the surrounding context words. The number of words that will be used as input is determined by the value of n or commonly referred to as n_window. When n is setted as 2, the model will predict the target word based on 2 words before the target word and 2 words after the target word, as long as the words are contained in one sentence or document. w(t) is the target word to be predicted, and w(t-n) is the words before the target word and w(t+n) is the words after the target word as far as n that have been previously determined. There is a projection layer between the input layer and the output layer, where this projection layer is a hidden layer that will store the projection of the training results in the form of a vector representation. The projection layer is predefined with n dimensions between 100 and 500.

Skip Gram

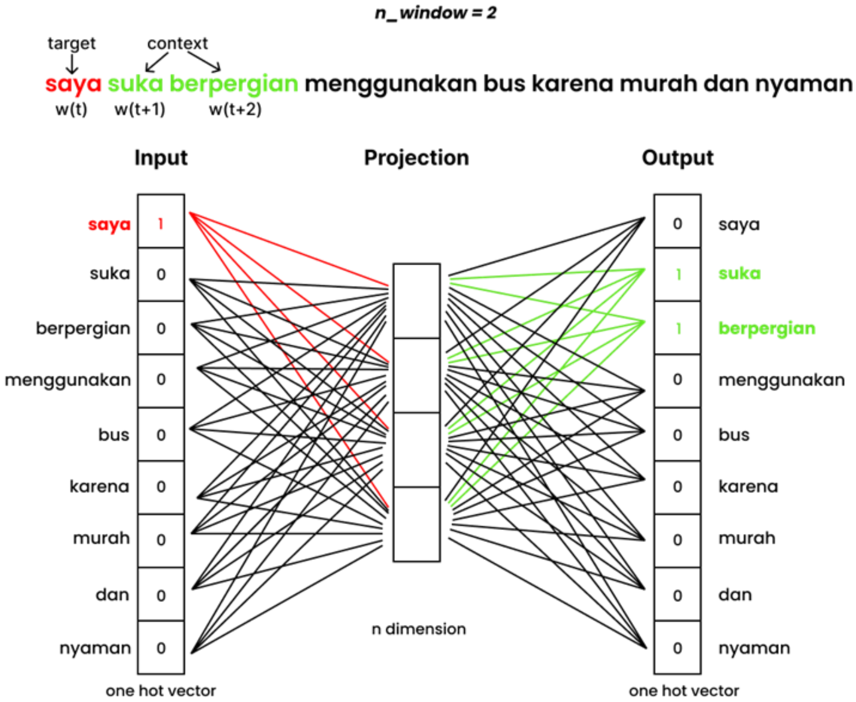

Figure 2. The Architecture of Skip-Gram

In the skip-gram architecture, Word2Vec does the opposite of Continuous Bag of Words. As depicted in 0. The fundamental difference between CBOW and Skip-gram is that Skip-gram will make the context word as the word to be predicted based on the target word specified as input. Just like CBOW, the context word is the word that is in the vicinity of the target word as far as the n window specified at the beginning. As shown in the 0, if the n window is set to 2, then the context word is w(t-2) to w(t+2) where w(t) is the target word. The result of Word2Vec model training is also stored into the projection layer with n dimensions that we specify at the beginning, the number of dimensions can range from 100 to 500.

CONVOLUTIONAL NEURAL NETWORK

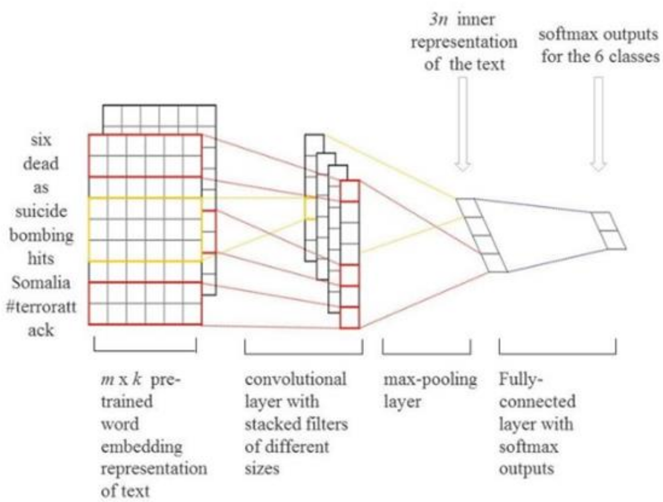

Figure 3. Convolutional Neural Network Architecture for Sentence Classification

Convolutional Neural Network is one of the artificial neural network architectures that is generally used to process data that has a grid structure, such as images. Convolutional Neural Network is effectively used for pattern detection tasks in data that has spatial relationships. With the development in the world of natural language processing, especially the method to convert words into vector representation, we can now use word embedding to perform the task. This has an impact on the use of CNN which is not only limited to classifying image data, but can be used for classifying text data. The CNN architecture for classifying sentences is not much different from the image classification task. In the figure 0, it is a cnn architecture for performing sentence classification. In the CNN architecture for sentence classification tasks, the input used is in the form of a matrix with a size of m x n where the value is a vector representation of each word contained in the sentence where the vector is obtained by the word embedding method, m is the number of words contained in the sentence and n is the dimensional size of the vector representation formed through word embedding. The dimension of the vector is the dimension determined at the time of creating the word embedding model.

Just as CNN is generally used for image classification, in the CNN architecture for text classification tasks there is also a feature extraction mechanism. Features extraction mechanism using a filter or kernel with a size determined by the n_gram parameter where the size determines how many words will be captured during the convolution process. A filter of size 2 is called a bigram where during the convolution process it captures the relationship between two consecutive words. In addition, there is a trigram or filter with size 3 where the filter is used to capture the relationship of three consecutive words. In addition to 2 and 3 the filter size is free to be determined according to the needs, the larger the size, the more features will be captured. The filter size in CNN for sentence classification generally has a size of n x dimension of word embedding representation.

In the convolution process, the vector representation of each word is calculated with the filter using the dot product method, where the results will be stored in a place called the pooling layer. The filters used are generally more than one with various sizes according to the task to be accomplished, and the values in the filters are determined randomly. After going through the convolution process, the convolution results stored in the pooling layer are entered into the max-pooling layer where each result stored in the pooling layer is taken as the maximum value. After each maximum value is stored in the max-pooling layer, the results are then combined into the fully connected layer to integrate all features to understand the entire context. In this fully connected layer, the input obtained from the max-pooling layer is calculated with weights using dot product calculations then added with bias, or commonly referred to as feedforward in neural networks.

TRAINING AND EVALUATION

The sentiment analysis model used in this research was created using the python programming language. There are two sentiment analysis models that will be created, both of which use the Convolutional Neural Network algorithm for the sentence classification process, the first difference is that the model is created using the vector representation generated by the CBOW architecture as the input layer of the CNN model, the second vector representation is created with the Skip-gram architecture. Furthermore, the two models are compared using accuracy, precision, recall, and F1 Score calculations. As a final result, the model will be given data on five bus companies to conduct sentiment analysis, then the sentiment analysis results from the five bus companies will be compared by researchers to find out which bus companies have the most dominant positive, negative, and neutral sentiments.

IMPLEMENTATION AND RESULT

Data Pre-Processing

	Nama Perusahaan Bus	Review	tokens
0	PO Rosalia Indah	pelayanannya baik dari cetak tiket sampai chec...	[pelayanannya, baik, dari, cetak, tiket, sampa...
1	PO Rosalia Indah	Kantor pusat atau Head Office nya bus Rosalia ...	[kantor, pusat, atau, head, office, nya, bus, ...
2	PO Rosalia Indah	rest area po rosin solo ini ramainya dari pagi...	[rest, area, po, rosin, solo, ini, ramainya, d...
3	PO Rosalia Indah	kantor pusat PO Rosalia Indah.sangat padat saa...	[kantor, pusat, rosalia, indah, sangat, padat,...
4	PO Rosalia Indah	First time tried double decker bus. Lumayan la...	[first, time, tried, double, decker, bus, luma...
...	...	...	...
4975	PO Kencana	Armadanya bagus sekarang, on time juga utk pem...	[armadanya, bagus, sekarang, on, time, juga, u...
4976	PO Kencana	Tidak on time, ada keterlambatan armada tp tid...	[tidak, on, time, ada, keterlambatan, armada, ...
4977	PO Kencana	Armada bagus bersih tapi kurangnyaman sopirnya...	[armada, bagus, bersih, tapi, kurangnyaman, so...
4978	PO Kencana	cukup on time, armada bersih, tempat duduk bag...	[cukup, on, time, armada, bersih, tempat, dudu...
4979	PO Kencana	Shuttle travel di Semarang, pelayanan ramah, b...	[shuttle, travel, di, semarang, pelayanan, ram...

4980 rows × 3 columns

Figure 4. Tokenization Results

To visualize the results of the create_tokens process, in 0 we can see that the tokens column is the result of tokenization for each sentence contained in the review column. The tokens are stored based on each sentence. After each token is successfully formed, the next process is to remove any numbers, punctuation marks, non-alphanumeric characters, stopwords, and finally the stemming process.

	review	tokens	tokens_clean
39	Bisnya berhenti dlu dsni pasti.. buat makan/Sn...	[bisnya, berhenti, dlu, dsni, pasti, buat, mak...	[bisnya, berhenti, dlu, dsni, pasti, buat, mak...
40	Sangat sangaaattt tidak disarankan menggunakan...	[sangat, sangaaattt, tidak, disarankan, menggu...	[sangat, sangaaattt, tidak, disarankan, menggu...
41	Naik SHS 396 rute Jogja-Jakarta tanggal 19 Feb...	[naik, shs, 396, rute, jogja, jakarta, tanggal...	[naik, shs, rute, jogja, jakarta, tanggal, feb...
42	Lumuayan bersih, cukup luas, lumayan lengkap. ...	[lumuayan, bersih, cukup, luas, lumayan, lengk...	[lumuayan, bersih, cukup, luas, lumayan, lengk...
43	Pengalaman naik bis paling nyaman dari mulai p...	[pengalaman, naik, bis, paling, nyaman, dari, ...	[pengalaman, naik, bis, paling, nyaman, dari, ...
44	Pelayanan ramah, tempat lumayan nyaman, unt...	[pelayanan, ramah, tempat, lumayan, nyaman, un...	[pelayanan, ramah, tempat, lumayan, nyaman, un...

Figure 5. Result of The Text Cleaning Process

Figure 5 shows the results of the text cleaning process, where this process is the process of deleting all punctuation marks, special characters, emojis contained in a collection of words or tokens. The review column is a collection of data in the corpus used to train the word2vec model and sentiment analysis. The tokens column is the result of tokenizing the review column before entering the text cleaning process and the tokens_clean column is the result of the text cleaning process.

	review	tokens	stopwords_removal_results
10	Salah satu perusahaan yang terbaik dalam membe...	[salah, satu, perusahaan, yang, terbaik, dalam...	[salah, satu, perusahaan, terbaik, memberikan,...
11	pusat tapi menurutku tempatnya sempit, makanan...	[pusat, tapi, menurutku, tempatnya, sempit, ma...	[pusat, menurutku, tempatnya, sempit, makanann...
12	Pelayanan ramah, tertib, rapi, bersih,higienis...	[pelayanan, ramah, tertib, rapi, bersih, higie...	[pelayanan, ramah, tertib, rapi, bersih, higie...
13	Pusatnya armada Rosalia berkumpul, bisa jadi s...	[pusatnya, armada, rosalia, berkumpul, bisa, j...	[pusatnya, armada, rosalia, berkumpul, jadi, s...
14	PO yang terbaik dengan pelayanan ramah..armada...	[po, yang, terbaik, dengan, pelayanan, ramah, ...	[po, terbaik, pelayanan, ramah, armadanya, san...
15	Tempat pembelian tiket yg dekat dengan rumah,p...	[tempat, pembelian, tiket, yg, dekat, dengan, ...	[tempat, pembelian, tiket, yg, dekat, rumah, p...
16	Di tiket tertulis berangkat pukul 14.00 dibela...	[di, tiket, tertulis, berangkat, pukul, dibela...	[tiket, tertulis, berangkat, pukul, dibela, in...
17	BUKAN BARU KALI INI SAJA KECEWA MA BIS INI, JA...	[bukan, baru, kali, ini, saja, kecewa, ma, bis...	[bukan, baru, kali, kecewa, ma, bis, jadwalnya...
18	didalam bis tempat kaki lega banget. udah dr j...	[didalam, bis, tempat, kaki, lega, banget, uda...	[didalam, bis, tempat, kaki, lega, banget, uda...
19	Semoga selalu memberikan pelayanan jasa yang t...	[semoga, selalu, memberikan, pelayanan, jasa, ...	[semoga, selalu, memberikan, pelayanan, jasa, ...

Figure 6. Result of The Stopwords Removal Process

After all numbers, punctuation, special characters, and emojis have been removed from the tokens. The next step is to remove any stopwords from the set of words, which are words that have no meaning. 0 shows the results of the process of removing stopwords or words that often appear in a sentence and generally have no meaning or if the word is removed it will not affect the meaning of a sentence. In the review column is a collection of data contained in the corpus, in the tokens column is a collection of tokens resulting from the text cleaning process previously carried out. The stopwords_removal_results column is the result of the stopwords removal process. We can compare the results in the stopwords_removal_results column with the tokens column.

	review	tokens	tokens_after_stemming
10	Salah satu perusahaan yang terbaik dalam membe...	[salah, satu, perusahaan, terbaik, memberikan,...	[salah, satu, usaha, baik, beri, layan, konsum...
11	pusat tapi menurutku tempatnya sempit, makanan...	[pusat, menurutku, tempatnya, sempit, makanann...	[pusat, turut, tempat, sempit, makan, lumayan,...
12	Pelayanan ramah, tertib, rapi, bersih,higienis...	[pelayanan, ramah, tertib, rapi, bersih, higie...	[layan, ramah, tertib, rapi, bersih, higienis,...
13	Pusatnya armada Rosalia berkumpul, bisa jadi s...	[pusatnya, armada, rosalia, berkumpul, jadi, s...	[pusat, armada, rosalia, kumpul, jadi, spot, t...
14	PO yang terbaik dengan pelayanan ramah..armada...	[po, terbaik, pelayanan, ramah, armadanya, san...	[po, baik, layan, ramah, armada, sangat, banya...
15	Tempat pembelian tiket yg dekat dengan rumah,p...	[tempat, pembelian, tiket, yg, dekat, rumah, p...	[tempat, belian, tiket, yg, dekat, rumah, pili...
16	Di tiket tertulis berangkat pukul 14.00 dibela...	[tiket, tertulis, berangkat, pukul, dibela, in...	[tiket, tulis, berangkat, pukul, bela, in, bur...
17	BUKAN BARU KALI INI SAJA KECEWA MA BIS INI, JA...	[bukan, baru, kali, kecewa, ma, bis, jadwalnya...	[bukan, baru, kali, kecewa, ma, bis, jadwal, m...
18	didalam bis tempat kaki lega banget. udah dr j...	[didalam, bis, tempat, kaki, lega, banget, uda...	[dalam, bis, tempat, kaki, lega, banget, udah,...
19	Semoga selalu memberikan pelayanan jasa yang t...	[semoga, selalu, memberikan, pelayanan, jasa, ...	[moga, selalu, beri, layan, jasa, baik, nyaman...

Figure 7. Result of Stemming Process

After the process of removing stopwords and cleaning tokens from punctuation, numbers, and special characters is done. The last step is to change each word in the tokens into its root word form. 0 is the result of the stemming process. In the tokens column are tokens that have gone through the text cleaning and stopwords removal processes. While the tokens_after_stemming column is a collection of tokens that have been successfully converted to the root word through the stemming process.

Word2vec Model Experiment Setup

In this study, researchers used two architectures of the Word2Vec algorithm, namely Continuous Bag of Words and Skip-Gram. From each architecture, researchers create four types of models for each architecture, the first type of model uses a window size of two so that the representation of the word to be learned by the model will see two words before and two words after, with an embedding size of 100. The second type of model uses 2 window sizes and an embedding size of 500. Then for the third type of model, researchers use a window size of four so that the model will learn words based on four words before and four words after and an embedding size of 100. The fourth type of model researchers made a model with four window sizes and an embedding size of 500. The four types of models that researchers made for the CBOW and Skip Gram architectures so that the total models to be formed were eight models. The first architecture is Skip Gram, researchers made a skip gram model with window size 2 and embedding size 100, a skip gram model with window size 2 and embedding size 500, a skip gram model with window size 4 and embedding size 100, and a skip gram model with window size 4 and embedding size 500. For CBOW architecture, researchers made a cbow model with window size 2 and embedding size 100, cbow model with window size 2 and embedding size 500, cbow model with window size 4 and embedding size 100, and finally cbow model with window size 4 and embedding size 500. The purpose of researchers making the eight word embedding models is to find out whether models with small window size and embedding size, and models with large window size and embedding size, have a significant effect on the resulting vector representation and can also affect the results of sentiment analysis.

Results

Table 2. Evaluation Table Model CNN with Skip Gram Window Size 2 Vector Size 100

Split Data	Accuracy	Precission	Recall	F1 Score
80	0.820410	0.410990	0.423440	0.414879
50	0.820119	0.540390	0.549780	0.537511
20	0.846645	0.590764	0.637941	0.607831

Table 3. Evaluation Table Model CNN with Skip Gram Window Size 2 Vector Size 500

Split Data	Accuracy	Precission	Recall	F1 Score
80	0.825739	0.422484	0.396753	0.392788
50	0.849957	0.588325	0.548610	0.556783
20	0.846645	0.583231	0.609297	0.594519

Table 4. Evaluation Table Model CNN with Skip Gram Window Size 4 Vector Size 100

Split Data	Accuracy	Precission	Recall	F1 Score
80	0.832667	0.436177	0.446385	0.438936
50	0.827792	0.535155	0.544161	0.536331
20	0.870075	0.649461	0.683775	0.664015

Table 5. Evaluation Table Model CNN with Skip Gram Window Size 4 Vector Size 500

Split Data	Accuracy	Precission	Recall	F1 Score
80	0.826272	0.419551	0.391889	0.386644
50	0.825234	0.513229	0.525169	0.517467
20	0.884984	0.683856	0.644443	0.657239

Table 6. Evaluation Table Model CNN with CBOW Window Size 2 Vector Size 100

Split Data	Accuracy	Precission	Recall	F1 Score
80	0.827871	0.559951	0.547818	0.545863
50	0.861040	0.626242	0.629848	0.627901
20	0.874334	0.588360	0.567367	0.569588

Table 7. Evaluation Table Model CNN with CBOW Window Size 2 Vector Size 500

Split Data	Accuracy	Precission	Recall	F1 Score
80	0.875833	0.661154	0.586626	0.604413
50	0.875533	0.634800	0.611674	0.620304
20	0.889244	0.735312	0.651209	0.671361

Table 8. Evaluation Table Model CNN with CBOW Window Size 4 Vector Size 100

Split Data	Accuracy	Precission	Recall	F1 Score
80	0.855049	0.546483	0.543298	0.542281
50	0.872975	0.622788	0.619280	0.619241
20	0.861555	0.605468	0.611989	0.606942

Table 9. Evaluation Table Model CNN with CBOW Window Size 4 Vector Size 500

Split Data	Accuracy	Precission	Recall	F1 Score
80	0.865707	0.586716	0.554835	0.561718
50	0.884058	0.652495	0.619122	0.631584
20	0.867945	0.635422	0.628828	0.628805

Discussion

Based on the above experiments, it is known that at 80% testing set comparison for vector word embedding generated by skip gram architecture often fails to detect neutral sentiment review data. Whether the window size is 2 or 4 and the vector size is 100 or 500, the model is unable to detect neutral labeled data. This is because the distribution of labels on the dataset used by researchers is unbalanced and has a large gap between data with majority labels and data with minority labels, especially data with neutral labels. Even so for the CNN model with matrix word embedding generated by the Continuous Bag of Words architecture, the model is able to detect neutral sentiment data both in the 80%, 50% and 20% testing set divisions. However, the limitation that is still a shortcoming of this research is that each model tends to be more able to detect positive sentiment data, and of course this is due to the number of labels that are not balanced with each other. For future research, researchers suggest using datasets with a balanced distribution between labels. In addition, to produce a better vector representation, a larger dataset is needed so that the word embedding model can learn more variations of words and sentences.

Conclusion

From the results of experiments conducted by researchers, it can be concluded that there are shortcomings in the results formed by the model, especially when the testing size is 80%, which means that the training data is only 20% of the total data. For some models, especially models generated with skip grams as vector word embedding, at a testing size of 80% it still cannot detect neutral sentiment data so that data that should be neutral sentiment or that has a neutral label cannot be detected as neutral sentiment data. For future research if readers want to do the same research or improve the shortcomings in this study, the researcher has several suggestions of things that need to be done, First, researchers recommend using a dataset with a balanced distribution between labels. In this study, researchers had 2291 data labeled positive, 767 data labeled negative, and 547 data labeled neutral. The imbalance between labels resulted in the model often failing to predict neutral data, which is often predicted as positive sentiment data.

Second, make sure that when labeling the data, it is necessary to define what criteria are the references for labeling the review data correctly so that the labeling results are consistent so that bias does not occur.

Third, to produce a better vector representation, a larger dataset is needed so that the word embedding model can learn more variations of words and sentences.

References

[1] [1]"Pemerintah Agar Beri Insentif Bagi Pengusaha Transportasi Umum," *Koran Jakarta*, 2020. https://koran-jakarta.com/pemerintah-agar-beri-insentif-bagi-pengusaha-transportasi-umum/

[2] B. Liu, *Sentiment analysis: Mining opinions, sentiments, and emotions*, 2nd ed. New York: Cambridge University Press, 2015. doi: 10.1017/CBO9781139084789.

[3] Y. Li and T. Yang, "Word Embedding for Understanding Natural Language: A Survey," *Stud. Big Data*, vol. 26, pp. 83–104, 2018, doi: 10.1007/978-3-319-53817-4_4.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 1st Int. Conf. Learn. Represent. ICLR 2013 - Work. Track Proc., no. October, 2013.

[5] H. woo An and N. Moon, "Design of recommendation system for tourist spot using sentiment analysis based on CNN-LSTM," J. Ambient Intell. Humaniz. Comput., 2019, doi: 10.1007/s12652-019-01521-w.

[6] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 8, no. 4, pp. 1–25, 2018, doi: 10.1002/widm.1253.

[7] V. Oktaviani, B. Warsito, H. Yasin, R. Santoso, and Suparti, "Sentiment analysis of e-commerce application in Traveloka data review on Google Play site using Naïve Bayes classifier and association method," J. Phys. Conf. Ser., vol. 1943, no. 1, 2021, doi: 10.1088/1742-6596/1943/1/012147.

[8] N. D. Wulandari, M. H. Z. Nuri, and L. Kurniasari, "Customers' Satisfaction and Preferences Using Sentiment Analysis on Traveloka: The Case of Yogyakarta Special Region Hotels," Proc. 1st UMGESHIC Int. Semin. Heal. Soc. Sci. Humanit. (UMGESHIC-ISHSSH 2020), vol. 585, pp. 407–416, 2021, doi: 10.2991/assehr.k.211020.058.

[9] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, "Recommendation systems: Principles, methods and evaluation," Egypt. Informatics J., vol. 16, no. 3, pp. 261–273, 2015, doi: 10.1016/j.eij.2015.06.005.

[10] J. Shen, Y. Wei, and Y. Yang, "Collaborative filtering recommendation algorithm based on two stages of similarity learning and its optimization," in IFAC Proceedings Volumes (IFAC-PapersOnline), 2013, vol. 13, no. PART 1, pp. 335–340. doi: 10.3182/20130708-3-CN-2036.00068.

[11] L. Jiang, Y. Cheng, L. Yang, J. Li, H. Yan, and X. Wang, "A trust-based collaborative filtering algorithm for E-commerce recommendation system," *J. Ambient Intell. Humaniz. Comput.*, vol. 10, no. 8, pp. 3023–3034, Aug. 2019, doi: 10.1007/s12652-018-0928-7.