Classification of Mobile Application User Ratings Based on Data from Google Play Store

Kiki Ahmad Baihaqi\textsuperscript{1,2*}, Eko Sediyono\textsuperscript{1}, Christine Dewi\textsuperscript{1}, Indrastanti R. Widiasari\textsuperscript{1}, and Ahmad Fauzi\textsuperscript{2}

\textsuperscript{1}University of Kristen Satya Wacana, Salatiga, Indonesia
\textsuperscript{2}University of Buana Perjuangan Karawang, Karawang, Indonesia

Abstract. This research is a comparison using 3 (three) algorithms, namely Logistic Regression, K-Nearest Neighbor, and Support Vector Machine in sentiment analysis about the JMO application, as the main means for participants in the employment social security program, which plays a crucial role in providing services that meet participants' needs well. This research aims to compare three different classification algorithms for sentiment analysis in the Jamsostek mobile application. The process involves several stages, including data collection (Crawling), word separation (tokenizing), normalization, removal of common words (Stopword), and word simplification (Stemming). After the processing stage, the data is labeled and classified using a comparison of three algorithms. The results of the 3 tweet category algorithms tend to be positive and negative. From the Logistic Regression algorithm, the accuracy level achieved was 84.78%, the precision was 87.24%, and the recall was 62.16%, then the Support Vector Machine algorithm achieved an accuracy level was 89.13%, the precision was 86.67%, and the recall of 76.88%, and the KNN algorithm produced an accuracy level of 88.59%, precision of 91.07%, and recall of 71.88%.

1 Introduction

Current rapid advances in technology have resulted in new innovations to create various mobile applications to provide convenience to the Indonesian people. The government responds to developments in technology and information by adopting policies for innovation in public services. One example of a public service institution initiated by the government is BPJS Employment, which has developed the Jamsostek Mobile application to improve service quality. BPJS Employment functions as a legal entity that provides social protection to all workers from certain social and economic risks. The Jamsostek Mobile (JMO) application is designed to make things easier for BPJS Jamsostek customers by providing online services. Some of the benefits of the JMO application include simulating old-age insurance, checking balances, and submitting claims [1].

* Corresponding author: kikiahmad@ubpkarawang.ac.id
In using the JMO Application, of course, several obstacles will arise related to the accessibility of information and the smooth implementation of the services provided. Therefore, it is important to obtain reviews from users regarding their satisfaction with the JMO application. These reviews can be used as evaluation material for companies to assess the quality of the application. The type of review desired is in the form of sentiment analysis or opinion mining, which is a computational science that includes an approach to understanding public opinion through indications revealed in the structure of the text (JMO journal). The text referred to here refers to unstructured data, and text mining or text analysis is the process of exploring and analyzing the data to get the essence of the text document, to produce significant results by certain objectives [1]. Sentiment analysis or opinion mining is the activity of extracting and evaluating individual opinions, attitudes, responses, or perceptions of an object, such as an event, topic, product, or service. Through sentiment analysis, information can be obtained regarding positive and negative sentiments originating from the public, and this information can be a basis for consideration in the decision-making process. The purpose of sentiment analysis is to evaluate the innovations introduced by BPJS Employment to the community so that BPJS Employment can make improvements to deficiencies that may exist in its application [2].

In this research, researchers compared machine learning algorithms, namely the Logistic Regression algorithm, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) against JMO application users. Evaluation of the performance of each algorithm is measured by calculating a matrix using the Confusion Matrix, which provides accuracy, precision, and recall values so that you can find out sentiment related to public reviews of JMO application users which can be used as material for evaluation and consideration, as well as producing algorithm comparisons, with a total of 1200 data. The latest data was obtained through web scraping from Google Play Store [3].

2 Method

2.1 Logistic Regression

Linear Regression, or linear regression, is the simplest form of regression because it involves only one independent variable, namely X. Linear regression operates by fitting a straight line to fit the data, also known as the process of "fit the line," to produce regression model with the minimum possible error rate. The basic concept of linear regression comes from the straight-line equation. With a simple formulation, it can be explained as follows [4].

2.2 Support Vector Machine

A highly desired and popular learning technique, Support Vector Machine (SVM), has a number of qualities that make it superior. SVM is based on a solid theoretical foundation and is known to provide more accurate classification results compared to most other algorithms in a wide range of applications. Several studies have revealed that SVM is often
considered the most accurate method for text classification. Apart from that, SVM is also often used in the context of sentiment classification [6].

$$y_i (x_i \cdot w + b) \geq 1(1 - \epsilon_i) \epsilon_i \geq 0; \sum \epsilon_i \leq C$$  \hspace{1cm} (1)

### 2.3 K-Nearest Neighbour

The KNN algorithm works by classifying new data using a number of nearby data as a reference. In this process, the algorithm determines new data classes based on their proximity to existing data. Then look for the closest one from the testing data and take the closeness [5].

$$d(x, y) = \sqrt{\sum (x1_i - y1_i)^2 + (x2_i - y2_i)^2 + \ldots}$$  \hspace{1cm} (2)

Note:

- $d(x,y)$ : distance x and y
- $x$ : data to be classified
- $y$ : ambient data
- $I$ : number of features

### 3 Results And Discussion

In this research, several machine learning algorithms are used to compare performance in sentiment classification, especially those that have good results from previous research, namely the Logistic Regression Algorithm, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). If in this case the results are good, then it can be used as a reference for the same case and different data. **Fig. 1** depicts the steps in comparative research on the classification algorithms applied.

**Fig. 1.** Step by step in the research

### 3.1 Collecting Data

In this research, data was taken from scraping results from Google Play store with the Python programming language and the latest comment data was taken. This data is in the form of reviews and ratings given by JMO application users. The amount of data collected was 1,200 data with the data collection time period being the latest assessment [7].
Fig. 2. Collecting data process

3.2 Pre-Processing

At this stage, the first step is processing raw data through the processes of case folding, cleansing, tokenizing, stop words, and stemming (Fig. 3 – Fig. 7). The following is a further explanation regarding the steps in pre-processing the data [8].

**Case Folding and Cleansing** This stage is to change the text in the dataset to lowercase and remove all unnecessary characters, such as symbols, URL links, emoticons, numbers.

**Tokenization** consists of cutting sentences into a list of words that form sentences separated by commas and spaces so that the result is individual words collected in the form of matrix data which will later be used in the weighting process.
Fig. 4. Process and result tokenizing

Stemming is the process of breaking down word forms into basic forms. For example, inaugurating becomes inaugurated.

Fig. 5. Processing stemming

Fig. 6. Processing negasi
**Normalize:** This process is carried out to change non-standard words into standard words.

```python
def convertToSlugWord(ulasan):
    kamusSlang = pd.read_csv("slangword.txt", names=["slangword"], header=None)
    pattern = re.compile(".*\b\(' + '\'.join(kamusSlang.keys()) + ')\b")
    content = []
    for kata in ulasan:
        filter_slang = pattern.sub(lambda x: kamusSlang[kata], kata.lower())
        if filter_slang.startswith('tidak_'):
            kata_depan = 'tidak_
            kata_belakang = kata[6]:
        else:
            content.append(filter_slang)
    ulasan = content
    return ulasan

ulasan["normalisasi"] = ulasan["negasi"].apply(convertToSlugWord)
ulasan["negasi", "normalisasi"]
ulasan["negasi", "normalisasi"].to_csv("normalisation.csv", index=False)
```

Fig. 7. Processing normalizes

### 3.3 Implementation of Machine Learning Models

**Logistic Regression** Classification is an algorithm for testing the probability of a dependent variable with an independent variable.

**Support Vector Machine SVM** is a classification algorithm that has been widely used in sentiment analysis. SVM will create a dividing line (hyperplane) which plays a role in separating classes with positive labels and classes with negative labels.

**K-Nearest Neighbor** KNN classification is a classification algorithm that groups test data based on its closest distance (similarity) to training data. The data grouping process refers to the number of k values used.

### 3.4 Testing

Logistic Regression Model Analysis The results of testing the Logistic Regression method can be seen in Fig. 8.

![Logistic Regression Model](image)

Fig. 8. Result Model Logistic Regression The results of the model show a mapping of true positives and true negatives and vice versa.
Analysis Model of Support Vector Machine, result of method Support Vector Machine can be seen at Fig. 9.

Fig. 9. Result of model Support Vector Machine (SVM)

4 Conclusion

In this research, we have compared the performance of 3 classification algorithms, namely Logistic Regression, Support Vector Machine, and K-Nearest Neighbor in conducting sentiment analysis about JMO application users. There is a difference in understanding among the public regarding the JMO application. Data was taken from web scraping on Google Playstore with the latest 1,200 data. The results are as follows. From the results obtained, this application received a positive response with a figure of 84%. With the algorithm results that can be seen in the presentation of the result. The highest accuracy value results are found in the Support Vector Machine (SVM) algorithm, amounting to 89.13%. The highest precision value results are found in the K-Nearest Neighbor (KNN) algorithm, amounting to 91.07%. The highest recall value results are found in the Support Vector Machine (SVM) algorithm, amounting to 76.43%.

References


