COMPARATION OF DECISION TREE MODEL AND SUPPORT VECTOR MACHINE IN SENTIMENT ANALYSIS OF REVIEW DATASET SAMSUNG SSD 850 EVO AT NEW EGG SHOP

Muhammad Fahmi Julianto1(*), Yesni Malau2, Wahyutama Fitri Hidayat3, Wawan Nugroho4, Fintri Indriyani5

Computer Science1, Electrical Engineering2, Software Engineering3, Information System4, Information Technology5
Universitas Bina Sarana Informatika
www.bsi.ac.id
fahmifj@bsi.ac.id, yesni.ymu@bsi.ac.id, revelationtama.wfh@bsi.ac.id, wawan.wgh@bsi.ac.id, fintri.fni@bsi.ac.id
(*) Corresponding Author

Abstract
The development of information technology is currently growing very rapidly, including the impact on the hardware used. This can be exemplified in the use of hard drives that are starting to switch to SSDs. The process of selecting an SSD product to be used cannot be separated from the sources of information found on the internet. Through the internet, every user can provide reviews, both positive and negative reviews. With the many reviews regarding the Samsung 850 Evo SSD on the NewEgg Store, the author uses it to be processed into information, which will have new knowledge. Based on that, the author makes research, in the form of opinion classification by analyzing sentiment through a text mining approach. In this study, two classification models were used, namely Decision Tree and Support Vector Machine. The results of this study are in the form of a comparison of the 2 models used based on the accuracy and AUC values. Based on research, the Support Vector Machine model is better than the Decision Tree model. This conclusion can be proven by the accuracy value of the Support Vector Machine model resulting in a value of 0.87 or 87% while the accuracy value of the Decision Tree model produces a value of 0.82 or 82%. In addition, the AUC value of the Support Vector Machine model produces a value of 0.87 and the Decision Tree model produces a value of 0.82 or it can be said that the AUC value of the Support Vector Machine model is better than the Decision Tree model.

Keywords: Sentiment Analysis, Decision Tree, Support Vector Machine

INTRODUCTION
As one of the largest E-Commerce in America, NewEgg is an online retail company that focuses on selling computer equipment, both hardware and software. Because NewEgg is in the sales field, they rely heavily on reviews of every product they sell. At this time the author discusses
reviews on the Samsung Evo SSD product. On the other hand, Solid State Drive (SSD) is a data storage device that uses a series of ICs as memory that is used to store data or information.(Fadjar Efendi Rasyid, 2016).

Product reviews are a form of conveying consumer opinions and sentiments towards a product online. Product reviews today have a very important role in influencing consumer interest in a product(Siringoringo & Jamaludin, 2019). Where reviews from buyers who have purchased the product will appear in the review column. At this stage, prospective buyers will be able to see reviews from previous buyers. This condition is generally called sentiment analysis and can determine the interest of the prospective buyer because the previous buyer's review column has given a perspective on the SSD. Sentiment Analysis is a study consisting of Natural Language processing, linguistic computing, and text analysis, so that it can assist in identifying the opinions of a product submitted by users, where the reviews are usually divided into two classes, namely positive and negative.(Fanissa, Fauzi, & Adinugroho, 2018). Detection of classification patterns in this review is an interesting object of research using text mining. Text mining has become an interesting field of research because of the large amount of text that exists on the web. Text mining is an important field in the context of data mining to find interesting patterns in textual data(Said A. Salloum, Ahmad Qasim Al Hamad & Shaalan, 2017). Text mining also involves all activities in finding information and other important data from various textual sources(Hashimi, Hafez, & Mathkour, 2015).

Based on previous research by regarding sentiment analysis conducted by(Irene, 2017) from the scenario test, it can be seen that the Support Vector Machine algorithm can be used for film review cases with an F1-Score value of 84.9%. In addition, other research conducted by(Mardiana, Syahreva, & Tuslaela, 2019) The test results with the confusion matrix obtained an accuracy value of 83% for Neural Network, 52% for K-Nearest Neighbor, 83% for Support Vector Machine, and 81% for Decision Tree. This study shows that the Support Vector Machine and Neural Network methods are best for classifying positive comments and negative related to franchising.

Sentiment analysis using text mining can help the process of understanding textual data extraction and processing to obtain information contained in an opinion sentence, this process aims to obtain interesting models and relationships and can be presented in large volumes of data. (Ronen Feldman, Bar-Ilan University, Israel, James Sanger, ABS Ventures, Boston, 2006). Based on the description of the background that has been described, the problems found are whether reviews can affect the selection of the Samsung Evo SSD product, the author will conduct an analysis that aims to perform sentiment analysis using the Decision Tree algorithm and Support Vector Machine to analyze sentiment problems related to review of the Samsung Evo SSD for sale at the NewEgg Store.

**RESEARCH METHODS**

This research is a sentiment analysis process to classify positive and negative user reviews on the sale of the Samsung Evo SSD at the NewEgg Store and determine the accuracy results using 2 methods, namely Decision Tree (DT) and SVM (Support Vector Machine).

**Research subject**

The subject of this research is to use positive and negative review data taken from the evaluation of Samsung Evo SSD sales at the NewEgg Store.

**Types of research**

This research is a type of experimental research. This method tests the success of the hypothesis and relates it to the research problem. This experimental model aims to classify sentiment analysis about reviews of Samsung Evo SSD sales at the NewEgg Store. Data collection to obtain the data source used is a public data collection method by taking data on the sales review of the Samsung Evo SSD at the NewEgg Store. Data collection on Samsung Evo SSD sales at the NewEgg Store is carried out with the help of a Python tool to collect review data.

**Time and Place of Research**

a. Problem identification and needs analysis  
   At this stage, a search for problems related to positive and negative reviews of Samsung Evo SSD sales is carried out at the NewEgg Store.

b. Data collection  
   Collecting the necessary data from the Samsung Evo SSD sales review data on the NewEgg Store.

c. Experiment  
   This stage determines the model used to enter training data into the model and tests using Python tools for the method used.

d. Implementation  
   Applying the proposed KNN, SVM and Naive Bayes methods to determine the accuracy of the predictions used by the user.
e. Evaluation
To measure whether the model that has been developed is successful or not, an evaluation is carried out. Evaluation is used to measure the accuracy achieved by the model.

f. Writing
Writing in the form of research reports is carried out simultaneously or in parallel with other steps to be effective and efficient.

Table 1 below shows that the research schedule is carried out for four weeks by carrying out activities that have been arranged according to the planned schedule. Table 1. Research Time

<table>
<thead>
<tr>
<th>N</th>
<th>ACTIVITY</th>
<th>WEEK</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Search and selection of research objects</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Study object of research</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>Research problem formulation</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>Topic determination</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>Reference material collection</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Preparation of the framework / rationale</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Preparation of research methodology/design</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Preparation of research proposal</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Manuscript</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Submission of research proposal</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Implementation of sample data collection</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Data analysis</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Preparation of the final research manuscript</td>
<td></td>
</tr>
</tbody>
</table>

Research Target / Subject
The purpose of this study is to analyze problems related to several reviews about the sale of SSD Samsung Evo at the NewEgg Store globally to determine whether it tends to be positive or negative and to show accuracy results based on the two methods used. In addition, it is also used to find the best method used by comparing the accuracy value and the AUC value.

Data, Instruments, and Data Collection Techniques
In this study, the dataset used is a review of the sale of the Samsung Evo SSD at the NewEgg Store which was obtained from the website https://www.kaggle.com/abdulrahmanalqamash/ssd-reviews which consists of 3108 data consisting of pros and cons labels.

Data analysis technique
The data analysis technique used is text mining using python programming where sentiment analysis is used to identify positive and negative opinions. In addition, a comparison of the best methods is proposed, namely Decision Tree and Support Vector Machine. Steps used:
1. Definition of dataset
   The dataset is defined by using pros and cons labels.
2. Pre-processing
   At this stage, two pros and cons labels are combined which are then labeled into positive and negative statements and then stopwords are removed.
3. Transformation
   The weighting of textual data, the process used is TF-IDF.
4. Classification
   Text classifiers usually use the Decision Tree and SVM methods.
5. Interpretation/Evaluation
   At this stage an evaluation is carried out to calculate the accuracy value and AUC value.

RESULTS AND DISCUSSION

1. Definition of Dataset
   The dataset used in this study is public data regarding sales of the Samsung Evo SSD at the NewEgg Store which consists of 10 data attributes. However, only 2 attributes are used in this study, namely pros which consists of 2205 data and cons which consists of 2216 data.

Based on Figure 1 above, it explains the number of words used in each label, but only pros and cons labels are used. To add a dataset and view the info, use the following code:

```python
df = pd.read_csv('ssd_reviews.csv', index_col=0)
df.info()
```

2. Pre-processing
   The pre-processing stage in data mining is to transform data into a format so that the process is
easier and more effective (Meilina, 2015). Preprocessing can be done in two conditions, the first is to form training data and the second is to detect intrusions (Jacobus & Winarko, 2014). In this study, pre-processing, the data that has been collected is first processed by labeling, and carrying out the stopwords removal process.

Figure 2. Data Pre-processing Process

Figure 2 describes the process of determining each review used in this study where 0 indicates that the review is negative while 1 indicates that the review is positive. The code used is as follows:

```python
df_cons = df['cons'].dropna()
df_cons['positive'] = 0
df_cons.drop(df_cons[df_cons['cons'].isin(['none', 'none so far', 'non'])].index, inplace=True)
df_cons.rename(columns={'cons': 'pros_and_cons'}, inplace=True)
df_pros = df['pros'][:1562].dropna()
df_pros.rename(columns={'pros': 'pros_and_cons'}, inplace=True)
df_pros['positive'] = 1
merged_df = pd.merge(left=df_pros, right=df_cons, on='pros_and_cons', how='outer')
df_cons = df[['cons']].dropna()
df_cons['positive'] = 0
df_cons.drop(df_cons[df_cons['cons'].isin(['none', 'none so far', 'non'])].index, inplace=True)
df_cons.rename(columns={'cons': 'pros_and_cons'}, inplace=True)
df_pros = df['pros'][:1562].dropna()
df_pros.rename(columns={'pros': 'pros_and_cons'}, inplace=True)
df_pros['positive'] = 1
merged_df = pd.merge(left=df_pros, right=df_cons, on='pros_and_cons', how='outer')
```

3. Transformation

The process of transforming data into a certain format so that the data is in accordance with the data mining process (June Arta, Indrawan, & Dantes, 2017). At this stage, TF-IDF is used in weighting the data that has been obtained. The weighting process with TF-IDF uses the following code:

```python
tfidf_vectoriser = TfidfVectorizer(stop_word='english')
tfidf_f = tfidf_vectoriser.fit(x['pros_and_cons'])
tfidf_transform = tfidf_f.transform(x['pros_and_cons'])
```

4. Classification Decision Tree

The first classification process used the Decision Tree model. The reason for choosing the decision tree model is that this algorithm produces a model that can predict data categories by determining categories based on data features (Ceballos, 2019) (Ochiai, Masuma, & Tomii, 2019). The following is the result of data processing using Decision Tree.

Figure 3. Value of x test and y train Decision Tree Model

For the x train and y train processes, the following code is used:

```python
Tree = DecisionTreeClassifier()
Tree.fit(tf_x_train, y_train)
print('test score', tree.score(tf_x_test, y_test))
y_pred = tree.predict(tf_x_test)
```

Figure 4. Classification Accuracy Value With Decision Tree Model

As for the classification process with the decision tree model, the following code is used:

```python
print(classification_report(y_test, y_pred))
print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred))
accuracy_entropy = metrics.accuracy_score(y_test, y_pred)
print("accuracy:" accuracy_entropy)
```

Based on Figure 3 and Figure 4 above, it can be stated that the classification results using the Decision Tree test model with x tran and y train yields a value of 0.99 while the test score using x test and y test produces a value of 0.82. The results of the accuracy value using the Decision Tree model produce a value of 0.82. The resulting confusion matrix value is shown in the following figure:
To display the confusion matrix of the decision tree model classification process, the following code is used:

```python
Print("Confusion Matrix :

sns.heatmap(confusion_matrix(y_test, y_pred, normalize='true'). Annot = True))
```

Meanwhile, to display the ROC AUC score of the decision tree model, the following code is used:

```python
def_multiclass_roc_auc_score(y_test, y_pred, average="macro"):
    lb = preprocessing.LabelBinarizer()
    lb.fit(y_test)
    y_test = lb.transforms(y_test)
    y_pred = lb_transform(y_pred)
    return roc_auc_score(y_test, y_pred, average=average)

print(ROC AUC score,’
multiclass_roc_auc_score(y_test,y_pred)
```

Based on Figure 5 and Figure 6 the confusion matrix Decision Tree model produces a true positive value of 0.84, a false negative of 0.16, a false positive of 0.18, and a true negative of 0.82. Meanwhile, the Area Under Curve (AUC) score is 0.82.

5. Classification Support Vector Machine

The second classification process used the Support Vector Machine model. The reason for using the SVM model is that this model uses a kernel function to determine the feature space and where the classifier function will be searched (Parapat, Furqon, & Sutrisno, 2018). In addition, SVM uses a hypothetical space in the form of linear functions in a high-dimensional feature and is trained using a learning algorithm based on optimization theory (Puspitasari, Ratnawati, & Widodo, 2018). The following is the result of data processing using the Support Vector Machine:

```python
Figure 7. Value of x test and y train Model Support Vector Machine
```

For the x train and y train processes, the following code is used:

```python
svm_linear = svm.SVC(kernel='linear')
svm_linear.fit(tf_x_train, y_train)
print(’Train ’, svm_linear.score(tf_x_train, y_train))
print(’Test ’, svm_linear.score(tf_x_test, y_test))
y_pred = svm_linear.predict(tf_x_test)
```

Based on Figure 7 and Figure 8 above, it can be stated that the classification results using the Support Vector Machine test score model with x tran and y train yields a value of 0.95 while the test score using x test and y test produces a value of 0.87. The results of the accuracy value using the Decision Tree model produce a value of 0.87. The resulting confusion matrix value is shown in the following figure:
To display the confusion matrix of the SVM model classification process, the following code is used:

```python
print("Confusion Matrix:

sns.heatmap(confusion_matrix(y_test, y_pred, normalize='true'). Annot = True))
```

Meanwhile, to display the ROC AUC score of the decision tree model, the following code is used:

```python
def_multiclass_roc_auc_score(y_test, y_pred, average="macro"):
    lb = preprocessing.LabelBinarizer()
    lb.fit(y_test)
    y_test = lb.transform(y_test)
    y_pred = lb.transform(y_pred)
    return roc_auc_score(y_test, y_pred, average=average)
print('ROC AUC score:', multiclass_roc_auc_score(y_test,y_pred)
```

Based on Figure 9 and Figure 10 the confusion matrix Support Vector Machine model produces a true positive value of 0.89, a false negative of 0.11, a false positive of 0.14, and a true negative of 0.86. Meanwhile, the Area Under Curve (AUC) score is 0.87.

6. Interpretation/Evaluation

The interpretation and evaluation stages are used to compare the accuracy results and the AUC 2 models used are Decision Tree and Support Vector Machine.

**CONCLUSIONS AND SUGGESTIONS**

**Conclusion**

Based on the discussion and research objectives, it can be concluded that the sales dataset of the Samsung 850 Evo SSD at the NewEgg Store with the classification of the Support Vector Machine model is better than the decision tree model. This conclusion can be proven by the accuracy value of the Support Vector Machine model resulting in a value of 0.87 or 87% while the accuracy value of the Decision Tree model produces a value of 0.82 or 82%. In addition, the AUC value of the Support Vector Machine model produces a value of 0.87 and the Decision Tree model produces a value of 0.82 or it can be said that the AUC value of the Support Vector Machine model is better than the Decision Tree model.

**Suggestion**

Suggestions for further research, this research can be used as a reference or previous research and use a classification model other than the one already used. So that later it can be used as a comparison of research that has been done.

**REFERENCES**


Juni Arta, IK, Indrawan, G., & Dantes, GR (2017). Data Mining Recommendations for Outstanding Student Candidates at STMIC Denpasar Using the Technique for Others Reference By Similarity To Ideal Solution Method. JST (Journal of Science and Technology), 5(2), 792. https://doi.org/10.23887/jst-undiksha.v5i2.8549


