International Journal of Engineering, Science and InformationTechnology
Volume 4, No. 3 (2024) pp. 118-127
ISSN 2775-2674 (online)
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v4i3.543
Research Paper, Short Communication, Review, Technical Paper

Analysis of Public Sentiment Toward Celebrity Endorsement On Media Social Using Support Vector Machine

M Oriza Syahputra*, Bustami, Lidya Rosnita

Department of Informatics, Faculty of Engineering, Universitas Malikussaleh, Aceh, Indonesia

**Corresponding author E-mail: oriza.180170087@mhs.unimal.ac.id*

The manuscript was received on 1 March 2024, revised on 15 April 2024, and accepted on 10 September 2024, date of publication 22 September 2024

Abstract

Analysis of public sentiment towards celebrity endorsements on social media is very important to understand the public's response to promotional campaigns involving celebrities. In this study, we combine the VADER labeling method with the Support Vector Machine (SVM) method to analyze public sentiment toward celebrity endorsements on social media. Data is taken from various social media sources such as Twitter, Instagram, and Facebook. The data is pre-processed to ensure data accuracy and relevance and then labeled with the VADER method to determine the positive, negative, or neutral sentiment of the text. The labeled data is then extracted for features and used to train the SVM model. The trained SVM model is then validated using test data to measure its accuracy and performance. The results of the analysis provide useful insight into public sentiment towards celebrity endorsements on social media and can provide recommendations for stakeholders regarding this matter. Overall, combining the VADER labeling method with SVM in analyzing public sentiment towards celebrity endorsements on social media shows more accurate results and can provide practical benefits in marketing and promotional strategies. The results shown using the Support Vector Machine method with a ratio of 80:20 can provide average precision results of 77%, recall of 100%, f1-score of 87%, and accuracy of 76.92%. Twitter application user sentiment shows that 77% (338 data) of Twitter user reviews provide positive sentiment and 23% (119 data) provide negative sentiment reviews from a total of 517 data. Suggestions from researchers are that in future research they can add more data to make modeling easier to provide higher accuracy values. Using other classification and performance evaluation methods, such as Naive Bayes, Decision Tree, Fuzzy, or Deep Learning. Use other data processing tools, such as RapidMiner, Jupyter Notebook, RStudio, or others.

Keywords: *Celebrity Endorsement, SVM, VADER, Deep Learning, Social Media.*

1. Introduction

Current developments in science and technology have an impact on various aspects of life, including the business world. As time goes by, many companies are starting to shift from traditional sales systems to more modern systems, especially with the emergence of new shopping platforms such as social media. This development encourages companies to market products online, giving consumers more choices and convenience in transactions. This change also requires business actors to meet customer needs by providing quality products, competitive prices, and adequate customer service [1].

Social media has various types, such as Facebook, Twitter, Telegram, Line, WhatsApp, and Instagram. Instagram itself was launched in 2010 to allow users to share photos and videos. As time goes by, Instagram continues to experience developments in its features. At first, Instagram only allowed users to post photos and videos, equipped with editing features for cropping and adding filters. Currently, Instagram has added various new features such as IG TV, Boomerang, Rewind, SuperZoom, Face Filter, Hashtags, Stickers, Direct Messages, Live Videos, InstaStory which can be saved in highlights, and many more again [2].

Advertising is one of the important elements in the marketing strategy used by companies. Effective advertising is advertising that can attract market attention. The Luwak White Coffee coffee product became popular thanks to a promotional campaign that utilized celebrity endorsements to convey the product message and build a positive image. The company chose Korean celebrities who have high popularity in the world to attract consumer attention. By using celebrities for promotions, companies hope to encourage buyers to buy the products they create [3].

Celebrity endorsement is the use of someone who has a reputation and achievements in a particular field and is widely known by the public, such as entertainers, actors/actresses, athletes, etc. Product endorsements can be carried out by various types of endorsers, including new endorsers, experts, or celebrities. When businesses use celebrities as endorsers, consumers tend to become more selective in choosing products. This can also increase the status of consumers because they own or use the same products as those promoted by the celebrity [4].

Sentiment analysis, known as opinion mining, is part of natural language processing used to identify and understand people's feelings or views on a particular product or topic. This analysis focuses on tracking and measuring public opinion. In this research, the author will

conduct research on public sentiment analysis towards celebrity endorsements on social media using the Support Vector Machine method [5].

2. Literature Review

2.1. Sentiment Analysis

Sentiment analysis known as opinion mining, is part of data mining and is also often used as a means of analyzing text data, understanding, processing, and retrieving textual data that contains opinions about certain entities, such as products, services, organizations, individuals, and topics. Specific. This analysis is used as a means of extracting certain information from the available data. Sentiment analysis is part of research in the field of Text Mining. It focuses on the computational study of opinions, feelings, and emotions expressed through text. The goal of sentiment analysis is to extract attributes from a review, such as opinions, feelings, and emotions expressed in writing. Once these attributes are extracted, the comments are evaluated to determine whether they have positive or negative sentiments [6]. Sentiment analysis is a field of study that studies opinions, attitudes, judgments, and evaluations of events, topics, organizations, and individuals[7]. Sentiment analysis is a method used to retrieve opinion data and understand and process text data automatically to identify the sentiment contained in an opinion[8].

2.2. Celebrity Endorsement

Celebrity endorsement is the use of famous people or public figures to support and promote an advertisement [9]. The selection of endorsers is related to brand attractiveness, namely the process of giving meaning or image to a brand. Celebrities will be more effective if they represent the main attributes they want to convey. The credibility of the advertising star also plays a very important role. Messages from highly credible sources tend to be more convincing. The three factors that determine credibility are expertise, trustworthiness, and attractiveness. Based on these concepts, it can be concluded that Celebrity Endorsement is a form of advertising that uses public figures as messengers to better communicate products, especially brands, to consumers [10].

2.3. Social Media

Social media with all its advantages has become an essential component in human life. Over time, various types of media have emerged, including social media. Social media is an online platform that provides its users with services to represent themselves, communicate, collaborate, share, interact with others, and build social relationships virtually. Social media functions as a digital space where social reality occurs and a place where users can speak in different contexts of space and time. Values in society or communities can appear in similar or different ways on the internet. According to several experts who study the internet, social media in cyberspace reflects events in the real world, including phenomena such as plagiarism [11].

2.4. Text Mining

Text Mining is the stage of extracting information from large text collections, and no longer using manual methods to identify patterns and significant relationships in text data. Text Mining is an interdisciplinary research, covering various fields such as data, natural language processing, machine learning, and information retrieval[12]. Text mining is used for various purposes, such as clustering, classification, information retrieval, and information extraction. Text Mining is a method used to solve problems such as classification, clustering, information extraction, and information retrieval. In general, the Text Mining process involves three main steps: text preprocessing, text mining, and post-processing. Text preprocessing tasks include data selection, classification, and feature extraction to transform documents into intermediate formats suitable for various search purposes. The main part of text mining operations involves clustering, association rule discovery, trend analysis, pattern identification, and application of knowledge discovery algorithms. The next step before processing is the manipulation of data or current information obtained from the text mining process, including evaluation and selection of information found, as well as interpretation and visualization of the results[13].

2.5. Classification

Classification is the process of determining a series of methods or functions to explain and distinguish categories of datasets. The main purpose of classification is to predict the category of data whose category has not been found. Classification involves two main processes: first, measuring the classification model based on a set of previously defined class data (training data), second, using the model or method to classify test data (prediction) can also measure the accuracy of the model. Classification can also be applied in various fields, including medical diagnosis, selective marketing, bank credit assessment, email filtering, and opinion analysis. Various classification models that are also used include *decision trees, Support Vector Machine classifiers, neural networks, k-nearest neighbors classifiers, and Naïve Bayes*.

2.6. VADER

VADER is commonly used as a sentiment analysis model to identify data variations based on the intensity of sentiment contained in the lexicon dictionary[14]. Sentences, sentiments, and phrases are evaluated using the lexicon dictionary to determine the polarity value of a word. The process of determining the polarity value of a sentence using the VADER method can be seen in Figure 1.

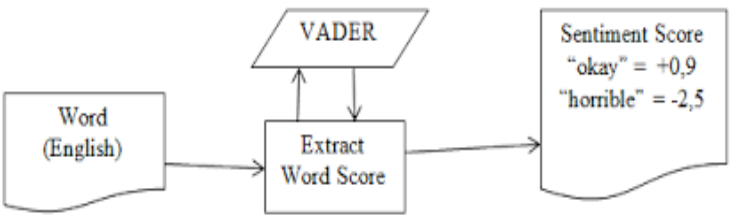

Fig 1. Groove of the value of polarity

2.7. TF-IDF

The TF-IDF method is a common technique used to calculate the weight of each word in information retrieval. This method is known for its effectiveness, ease of use, and accurate results. In this method, the Term Frequency (TF) and Inverse Document Frequency (IDF) values are calculated for each word in each document in the corpus. The process is through calculating the weight of each word in the document using the formula:

$$tf = 0{,}5 + 0{,}5 * \frac{ft{,}d}{\max(ft{,}d)} \quad (1)$$

Term Frequency (TF) calculates the frequency of occurrence of a term in a document. Document Frequency (DF) calculates the number of data containing the term. As seen in the formula above, this calculation involves the value of term frequency and document frequency. Where N is the total number of documents indicating the document frequency for the term being searched for inversely. Inverse Document Frequency (IDF) is calculated using the following formula:

$$IDF = Log\ \frac{n}{df} \quad (2)$$

$$W = tf * idf$$

Once the weight (W) for each document is determined, a sorting process is performed, where documents with higher W values indicate a greater degree of similarity to the keyword, while lower W values indicate a lesser similarity.
Information:
d: document -d
t: the tth word of the keyword
W: the weight of the dth document to the tth word
TF: The number of words in the document being searched
IDF: Inverse Document Frequency
ft,d: the frequency of words in d
df: many documents containing search words

2.8. Support Vector Machine

Support Vector Machine (SVM) classification is one of the machine learning methods that utilizes linear functions in high-dimensional feature spaces. This system also uses an algorithm based on the concept of optimization, by applying study deviations obtained from statistical study theory. Support Vector Machine (SVM) is a modern scalable prediction method for classification and regression tasks. SVM falls into the category of supervised studies, where it requires a training phase followed by a testing phase. This process involves gradual training with SVM before the testing phase is carried out[15].
Support Vector Machine (SVM) is a classification method in machine learning (supervised learning) that is used to predict classes regarding shapes or patterns based on training results. The assessment is processed by determining a hyperplane or can be called a decision boundary that separates one class from another, in this context separating positive sentiment (marked with +1) and negative sentiment (marked with -1). SVM determines the hyperplane by utilizing support vectors and margins. In this study, the input data is represented as a vector obtained through a weighing process. After training in the SVM classification, the values and patterns used for testing the SVM are generated to mark reviews in tweets. The process stages in the classification problem aim to find a line (or hyperplane) to separate two groups. The picture shows the stages of the study in SVM:

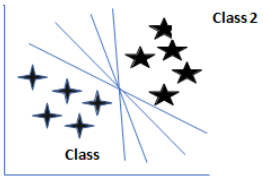

Fig 2. *Hyperplane* Positive and Negative Classes

3. Methods

Some methodologies used to obtain data or information to solve problems include:

1. Literature Study
   The data collection stage through literature involves collecting journals, literature, papers, books, and sources from internet sites that are relevant to the topic, especially regarding sentiment analysis with Support Vector Machine (SVM) classification.
2. Celebrity Endorsement *Data Collection*
   Data was collected directly from social media Twitter and Instagram using *Tweet Harvest* with the search keyword #product.
3. Planning
   At this stage, the steps taken include: designing the data collection flow, designing the system and method flow, and designing the program.
4. Testing and Evaluation
   At this stage, a testing process will be carried out on the system to identify any potential errors that could occur and find solutions to overcome them.
5. Preparation of Reports
   The report is prepared as documentation to facilitate understanding and development by other parties in the future

4. Results and Discussion

4.1. Research Result

In this study, the author uses the *Term Frequency -Inverse Document Frequency* (TF-IDF) extraction method and *Support Vector Machine* (SVM) modeling to identify reviews of *celebrity endorsement* on social media, both positive and negative. The *Term Frequency - Inverse Document Frequency* (TF-IDF) method is used to classify *terms* in a *document* correctly according to *the positive or negative* class.

4.2. System Analysis

System analysis is a research process to evaluate whether the procedures or systems that have been implemented by a company are following standards to improve efficiency. Based on the Great Dictionary of the Indonesian Language (KBBI), system analysis refers to a systematic process that supports a combination of considerations from experts in a particular field, to achieve optimal results from each function of the discipline applied. It also includes observations of certain activities, methods, procedures, or ways to determine the benefits of these activities, as well as the best techniques in their implementation.

System analysis can be understood as a technique for solving problems by breaking down a system into its constituent components. The goal of this process is to evaluate the performance of each component and to understand how the interactions between these components contribute to the achievement of the overall system goals. The system to be built is a public sentiment analysis system for *celebrity endorsements* on social media using the *Support Vector Machine (SVM)* method which produces *output* in the form of a *class*, namely positive or negative.

4.3. Crawling Data Using *TweetHarvest*

Crawling review data on the Twitter application using the Tweet Harvest library using the Google Colab tools. The data taken uses *the keyword "celebrity endorsement"*. The data was obtained from January 1, 2017, to March 28, 2024. Furthermore, the crawled data will be used in *the support vector machine* modeling process. Here are the stages of data crawling and the script:

1. Install and import *the* pandas' library that works for data manipulation and analysis, and import *json library* to process data in JSON format.

```
# Import required Python package
!pip install pandas
# Install Node.js (because tweet-harvest built using Node.js)
!sudo apt-get update
!sudo apt-get install -y ca-certificates curl gnupg
!sudo mkdir -p /etc/apt/keyrings
!curl-fsSL https://deb.nodesource.com/gpgkey/nodesource-repo.gpg.key | sudo gpg --dearmor -
o /etc/apt/keyrings/nodesource.gpg
!NODE_MAJOR=20&&echo"deb[signed-by=/etc/apt/keyrings/nodesource.gpg]
https://deb.nodesource.com/node_$NODE_MAJOR.x nodistro main" | sudo tee
/etc/apt/sources.list.d/nodesource.list
!sudo apt-get update
!sudo apt-get install nodejs -y
!node -v
```

2. Input *keywords* for the data mining process and input time to determine since and equal when the data starts to be taken

```
# Crawl Data
filename = 'Crawl2021.csv'
search_keyword = 'Celebrity Endorses until:2017-12-31 since:2021-01-01'
limit = 1000
!npx --yes tweet-harvest@latest -o "{filename}" -s "{search_keyword}" -l {limit} --token ""
```

Here are 10 data that have been taken from 517 data that have been crawled.

Table 1. Review Data

No	Ulasan
1	*Benefits of being a civil servant (Nagita Slavina Employee) - Salary is certain. - Opportunity to be invited abroad - Get branded goods from Raffi & Gigi - Often meet Capital Artists, you can build relationships with them - Followers increase, so celebrities get endorsements*
2	*Is it true that Online Game Endorsing Artists Make Real Money??? This is an online game but makes real money. Isn't this called online gambling?? I thought that gambling sites had fallen into disrepair, but they are increasingly mushrooming. Ouch??*
3	*.. Artists who endorse Fuckcuin get door prize geis !! Right on target !! .. Previous Hammer kiwil Indr4B3kti etc.*
4	*Artists who just appeared in PS STORE*
5	*-Cinema taxes go up -People normalize pirated nnton -The film industry is sluggish -Demo artists take to the streets -Trus is beaten by the masses because they used to endorse the creation*
6	*Sump! Where can I get to eat as GABRUG as this 29k guys. Impossible! Believe me, the Hajj Endors, that this is ENDORSED!! Today, if you don't endorse an artist, it's not fun!*
7	*I want to be an artist, please if you want to endorse it, you can contact me.*
8	*Still in the edition of Thai artists like this Indonesian celebrity.*
9	*Can an artist who has been scandalized be banned from the entertainment world, no matter how small the scandal, if you have been scandalized, you can be directly banned, so don't have endorsements to enter this offer, even be interrogated or what else?*

10	*The type of person who is very stupid with the popularity of the artist usually exists more on ig for promotional purposes. JK made a bold decision, imagine if JK receives endorsements with the number of followers of 53 million 1 post, you can get 50 billion won or the equivalent of 584 billion rupiah*

1. Labeling on each comment will be done automatically using the nltk vader lexicon feature

```
# Install and import nltk
!pip install nltk
import nltk
# Download the lexicon
nltk.download("vader_lexicon")
# Import the lexicon
from nltk.sentiment.vader import SentimentIntensityAnalyzer
# Create an instance of SentimentIntensityAnalyzer
sent_analyzer = SentimentIntensityAnalyzer()
import pandas as pd
sentiment_data = pd.read_csv('dataset.csv', index_col=0)
sentiment_data.head()
def format_data(data):
  last_col = str(data.columns[-1])
  data.rename(columns = {last_col: 'ulasan'}, inplace=True)
  return data[['ulasan']]
data = format_data(sentiment_data)
data.head()
def format_output(output_dict):
  polarity = 1
  if(output_dict['compound']>= 0.05):
    polarity = 1
  elif(output_dict['compound']<= -0.05):
    polarity = 0
  return polarity
def predict_sentiment(text):
  output_dict =  sent_analyzer.polarity_scores(text)
  return format_output(output_dict)
data["vader_prediction"] = data["ulasan"].apply(predict_sentiment)
```

Here are 10 data results from labeling using vader:

Table 2. VADER Labeling Result

No	Ulasan	Sentimen
1	*Benefits of being a civil servant (Nagita Slavina Employee) - Salary is certain. - Opportunity to be invited abroad - Get branded goods from Raffi & Gigi - Often meet Capital Artists, you can build relationships with them - Followers increase, so celebrities get endorsements*	Positif
2	*Is it true that Online Game Endorsing Artists Make Real Money??? This is an online game but makes real money. Isn't this called online gambling?? I thought that gambling sites had fallen into disrepair, but they are increasingly mush-rooming. Ouch??*	negative
3	*.. Artists who endorse Fuckcuin get door prize geis !! Right on target !! .. Previous Hammer kiwil Indr4B3kti etc.*	Positif
4	*Artists who just appeared in PS STORE*	Positif
5	*-Cinema taxes go up -People normalize pirated nnton -The film industry is sluggish -Demo artists take to the streets -Trus is beaten by the masses because they used to endorse the creation*	Positif
6	*Sump! Where can I get to eat as GABRUG as this 29k guys. Imposible! Believe me, the Hajj Endors, that this is definitely ENDORSED!! Today, if you don't endorse an artist, it's not fun!*	Negatives
7	*I want to be an artist, please if you want to endorse it, you can contact me*	Positif
8	*Still in the edition of Thai artists like this Indonesian celebrity.*	Positif
9	*Can an artist who has been scandalized be banned from the entertainment world, no matter how small the scandal, if you have been scandalized, you can be directly banned, so don't have endorsements to enter this offer, even be interrogated or what else?*	Positif
10	*The type of person who is very stupid with the popularity of the artist usually exists more on ig for promotional purposes. JK really made a bold decision, imagine if JK receives endorsements with the number of followers of 53 million 1 post, you can get 50 billion won or the equivalent of 584 billion rupiah*	Negatif

4.4. Data Processing

Next, after the data is obtained and labeled using Vader, the data is processed through five data processes including cleansing, normalization, stopword removal, stemming, and tokenization. Before conducting sentiment analysis on the tweet data that has been collected, the data processing process needs to be carried out so that it is ready for analysis. Raw tweets obtained from Twitter must be processed first before being used in the analysis. This process begins with tweet cleaning and continues until it produces terms that will be given weighting. Here are the stages:

1. *Cleaning*
   This step will convert the text to lowercase, and clean the review column from special characters, punctuation, and irrelevant symbols. The following are the steps and scripts in the cleansing process
   a. Imports *the pandas* library to analyze the data, *regular expressions* (re) to search for data and replace it, *strings* to import punctuation characters, and *DataFrame* (df) to read *files.*
   b. Executes functions to convert text to lowercase, remove *non-alphanumeric characters*, text from *tabs*, *new lines*, *links*, *hashtags*, URLs, *back slice*, numbers, *non-ASCII*, *punctuation*, *and* excessive whitespace
   c. Run the function to clean reviews in the *"*reviews*"* column and save them in a new "*cleaning" column*

```
import pandas as pd
import re
import string
import json
import numpy as np
from tqdm import tqdm
def cleaningulasan(ulasan):
  ulasan = re.sub(r'@[A-Za-a0-9]+',' ',ulasan)
  ulasan = re.sub(r'#[A-Za-z0-9]+',' ',ulasan)
  ulasan = re.sub(r"http\S+",' ',ulasan)
  ulasan = re.sub(r'[0-9]+',' ',ulasan)
  ulasan = re.sub(r"[-()\"#/@;:<>{}'+=~|.!?,_]", " ", ulasan)
  ulasan = ulasan.strip(' ')
  return ulasan
data['Cleaning']= data['ulasan'].apply(cleaningulasan)
def clearEmoji(ulasan):
    return ulasan.encode('ascii', 'ignore').decode('ascii')
data['HapusEmoji']= data['Cleaning'].apply(clearEmoji)
```

The following are the results of the review before and after the *cleaning process* in the following table:

Table 3. Before and After Cleaning Review

No	*Cleaning*
1	*Benefits of being a civil servant Nagita Slavina Employee Salary is certain Opportunity to be invited abroad Get branded goods from Raffi &amp Gigi Often meet Capital Artists you can build relationships with them Followers increase so celebrities get endorsements.*
2	*Is it true that Online games endorsing Artists Make Real Money This is an online game but makes real money Isn't this called online gambling I thought that gambling sites had fallen into disrepair but they are increasingly mushrooming Ouch*
3	*Artists who endorse Fuckcuin get door prize geis Right on target Previous Hammer kiwil Indr B kti etc*
4	*Artists who just appeared in PS STORE*
5	*txtdarionlshop at the Korean Artist Guys endorsements*

2. *Normalization*
   The process goes on to convert slang words and abbreviations into their root words. Here are the stages and scripts in the *normalization process*
   a. Import *the pandas'* library (pd) to read and analyze the data.
   b. Execute a function to separate words and replace them with root words.
   c. Save it in the new *"normalized" column*. Then save the normalization results.

```
def normalization(Ulasan):
  normalized_text = Ulasan
  return normalized_text
data['normalized'] = data['Steming'].apply(normalization)
```

The following are the results of the review before and after *normalization* in the following table:

Table 4. Normalization Result

No	*Normalization*
1	*benefits civil servant nagita slavina employee salary certain opportunity invited abroad get branded goods raffi amp gigi often meet capital artists build relationships followers increase celebrities get endorsements*
2	*true online game endorsing artists make real money online game makes real money called online gambling thought gambling sites fallen disrepair increasingly mushrooming ouch*
3	*artists endorse fuckcuin get door prize geis right target previous hammer kiwil indr b kti etc*
4	*artists appeared ps store*
5	*txtdarionlshop korean artist guys endorsements*

3. *Stopword Removal*
   In this process, it will eliminate common words that do not provide important information. Here are the stages and *scripts* in the *stopwords removal process.*
   a. Import the *nltk library* for stopwords needs.
   b. Load and read *downloaded English* stopwords.
   c. Run a function to delete the data in the *"normalized" column* in the *file according to the word stopword* and save the result of the process in the new column *"stopword"*.

```
from nltk.corpus import stopwords
nltk.download('stopwords')
daftar_stopword = set(stopwords.words('english'))
def stopwordText(words):
 return [word for word in words if word not in daftar_stopword]
data['Stopword'] = data['Tokenizing'].apply(stopwordText)
```

Here are the results of the review before and after *stopwords* in the following table:

Table 5. Result Stopword

No	*Stopwords*
1	*benefits,civil,servant,nagita,slavina,employee,salary,certain,opportunity, invited, abroad,get,branded,goods,raffi,&amp,gigi,often,meet,capital,artists,build, relationships,followers,increase,celebrities,get,endorsements*
2	*true,online,game,endorsing,artists,make,real,money,online,game,makes, real,money,called,online,gambling,thought,gambling,sites,fallen,disrepair, increasingly,mushrooming,ouch*
3	*artists,endorse,fuckcuin,get,door,prize,geis,right,target,previous,hammer, kiwil,indr,b,kti,etc*
4	*artists,appeared,ps,store*
5	*txtdarionlshop,korean,artist,guys,endorsements*

4. *Steeming*
   The stemming process will reduce words that have suffixes into root words. Here are the stages and scripts in the *steaming process*
   a. *Install* the Literary dictionary and import the *StemmerFactory* module to stem United Kingdom texts.
   b. Running a function to remove meaningless or misspelled words will be removed and saved into a new "*stemming*" column

```
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory
def stemming(Ulasan):
  factory = StemmerFactory()
  stemmer = factory.create_stemmer()
  do = []
  for w in Ulasan:
    dt = stemmer.stem(w)
    do.append(dt)
  d_clean = []
  d_clean = " ".join(do)
  print(d_clean)
  return d_clean
data['Steming'] = data['Stopword'].apply(stemming)
```

Here are the results of the before and after stemming review in the following table :

Table 6. Result Steeming

No	*Stemming*
1	*benefits civil servant nagita slavina employee salary certain opportunity invited abroad get branded goods raffi amp gigi often meet capital artists build relationships followers increase celebrities get endorsements*
2	*true online game endorsing artists make real money online game makes real money called online gambling thought gambling sites fallen disrepair increasingly mushrooming ouch*
3	*artists endorse fuckcuin get door prize geis right target previous hammer kiwil indr b kti etc*
4	*artists appeared ps store*
5	*txtdarionlshop korean artist guys endorsements*

5. *Tokenization*
   The *tokenization* stage will break down sentences into words. Sentence breaking uses *unigram-bigram feature* extraction, which is breaking into one word and two words. Here are the stages and scripts in the *tokenization process*:
   a. Imports the *CountVectorizer* library to convert text into numerical representations.
   b. Execute a function to break down sentences in the "*stemming*" column into pieces of words. Next, save the tokenization result in the new "tokenization" column.

```
def tokenizingText(ulasan):
 ulasan = ulasan.split()
```

```
  return ulasan
data['Tokenizing']= data['CaseFolding'].apply(tokenizingText)
```

Here are the results of the before and after tokenization review in the following table:

Table 7. Result Tokenization

No	*Tokenization*
1	*bene-fits,of,being,a,civil,servant,nagita,slavina,employee,salary,is,certain,opportunity,to,be,invited,abroad,get,branded,goods,from,raffi,&amp,gigi,often,meet,capital,artists,you,can,build,relationships,with,them,followers,increase,so,celebrities,get,endorsements*
2	*is,it,true,that,online,game,endorsing,artists,make,real,money,this,is,an,online,game,but,makes,real,money,isn,t,this,called,online,gambling,i,thought,that,gambling,sites,had,fallen,into,disrepair,but,they,are,increasingly,mushrooming,ouch*
3	*art-ists,who,endorse,fuckcuin,get,door,prize,geis,right,on,target,previous,hammer,kiwil,indr,b,kti,etc*
4	*artists,who,just,appeared,in,ps,store*
5	*txtdarionlshop,at,the,korean,artist,guys,endorsements*

4.4. Data Modeling

This stage will generate accuracy values by conducting modeling training on positive and negative label data. The following are the stages and *scripts* of data modeling with the *Support Vector Machine* (SVM) classification method:

1. Importing the "*pandas*" library to read the data, *"train_test_split"* to separate the data into training data and test data, *"LabelEncoder"* to convert sentiment labels to numeric values, *"TfidfVectorizer"* to vectorize or weight the data, "SVC" *(Support Vector Classifier)* to perform classification with SVM, and *"accuracy_score"* to measure the value of prediction accuracy.
2. Sharing of training data and test data.
3. Weighting of review data in the "tokenization" column.
4. Carry out classification functions based on the division of training data and test data.
5. Performs accuracy functions.

```
from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer
Ulasan = data['normalized']
Ulasan.isnull().sum()
Ulasan = Ulasan.fillna('tidak ada komentar')
cv = CountVectorizer()
term_fit = cv.fit(Ulasan)
print(len(term_fit.vocabulary_))
term_fit.vocabulary_
term_frequency_all = term_fit.transform(Ulasan)
ulasan_tf = Ulasan[60]
term_frequency = term_fit.transform([ulasan_tf])
dokumen = term_fit.transform(Ulasan)
tfidf_transformer = TfidfTransformer().fit(dokumen)
print(tfidf_transformer.idf_)
tfidf = tfidf_transformer.transform(term_frequency)
print(tfidf)
from sklearn.model_selection import train_test_split
X_train,    X_test,    y_train,    y_test    =    train_test_split(data['normalized'],    da-
ta['vader_prediction'], test_size=0.1, stratify=data['vader_prediction'], random_state=30)
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(decode_error='replace', encoding='utf-8')
X_train = vectorizer.fit_transform(X_train)
X_test = vectorizer.transform(X_test)
print(X_train.shape)
print(X_test.shape)
X_train = X_train.toarray()
X_test = X_test.toarray()
from sklearn.svm import SVC
from sklearn.metrics import confusion_matrix, accuracy_score, classification_report
clf = SVC(kernel='linear')
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
CLF_acc = accuracy_score(y_pred, y_test)
print(classification_report(y_test, y_pred))
print("Akurasi SVM : {:.2f}%".format(CLF_acc*100))
```

	precision	recall	f1-score	support
0	0.00	0.00	0.00	12
1	0.77	1.00	0.87	40
accuracy			0.77	52
macro avg	0.38	0.50	0.43	52
weighted avg	0.59	0.77	0.67	52

Akurasi SVM : 76.92%

Fig 3. Accuracy Support Vector Machine (SVM)

4.5. Positive & Negative Sentiment Result Chart

The following is a graph of the results of the positive and negative sentiment research.

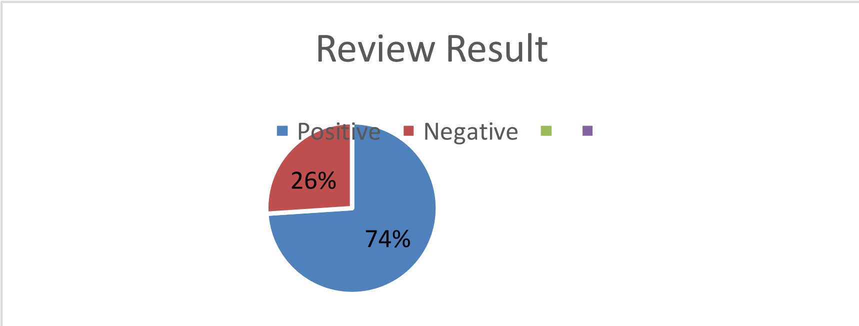

Fig 4. Result Graphic

4.6. Interpretation of Positive Sentiment Result

The interpretation of the results further shows the topics that are often talked about by Twitter users. Twitter user review topics will be displayed using word cloud visualizations and diagrams. The following shown in Figure 5 is a topic of discussion for users who provide positive sentiment

Fig 5. WordCloud Positive Review

4.7. Interpretation of Negative Sentiment Result

The interpretation of the results further shows the topics that are often talked about by Twitter users. Twitter review topics will be displayed using word cloud visualizations and diagrams. The following shown in Figure 6 is the topic of discussion of users who gave negative sentiments.

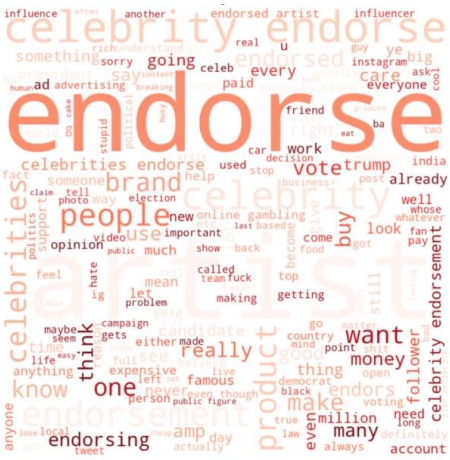

Fig 6. WordCloud Positive Review

5. Conclusion

Based on the results of the research that has been carried out, it can be concluded that:

1. The use of *unigram-bigram* feature extraction and data labeling using *a lexicon-based* approach can provide good sentiment labeling so that it can provide stable classification results on Twitter application user reviews. Twitter app user sentiment shows that 77% (338 data) of Twitter user reviews give positive sentiment and 23% (119 data) give negative sentiment reviews out of a total of 517 data.
2. The results shown through the 80:20 ratio of the Vector Machine support method can provide an average accuracy result of 77%, *recall of* 100%, *f1-score of* 87%, and accuracy of 76.92%.

References

[1] Ode, W., Nilammadi, M., Atin, T., Leksono, B. B., Sari, A. W., Syafiqatulbestariyah, A., Safitri, D. N., Akuntansi, J., Ekonomi, F., & Surabaya, U. N. (2020). Hubungan Celebrity Endorsement Terhadap Keputusan Pembeli Pada Shopee Dan Tokopedia. *Jurnal Bisnis dan Kajian Strategi Manajemen*. *5*, 1–13

[2] Watil, H. (2019). Pengaruh Selebgram sebagai Celebrity Endorsement terhadap Perilaku Konsumtif Mahasiswi *Jurnal Pendidikan Sosiologi*, *9*(1), 722–728. http://dx.doi.org/10.17509/ijost.v4i1.xxxx.

[3] Prasojo, E. (2019). Pengaruh Celebrity Endorsement Terhadap Minat Beli Melalui Brand Image (Studi Pada Mahasiswa Kota Tegal). *Journal Competency of Business*, *3*(1), 14–33. https://doi.org/10.47200/jcob.v3i1.667.

[4] Pradhan Debasis, Israel Duraipadian., and Dhruv Sethi. (2014). Celebrity Endorsement: How Celebrity- Brand – User Personality Congruence Affects Brand Attitude And Purchase Intention. Journal of Marketing Communications, 2(2): 1-18.

[5] Vinodhini, G., & Chandrasekaran, R. (2012). Sentiment Analysis and Opinion Mining : A Survey International Journal of Advanced Research in Sentiment Analysis and Opinion Mining : A Survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, *2*(6), 283–292.

[6] Pang, B., & Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. Journal Foundations and Trends in Information Retrieval *2*, 1–135. https://doi.org/10.1561/1500000001

[7] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, Chicago: Morgan & ClaypoolPublisher *5*(1), 1–184

[8] Sari, F. V., & Wibowo, A. (2019). Analisis Sentimen Pelanggan Toko Online Jd.Id Menggunakan Metode Naïve Bayes Classifier Berbasis Konversi Ikon Emosi. *Jurnal SIMETRIS*, *10*(2), 681–686

[9] A Shimp, Terence. (2007). *Periklanan Promosi (Aspek Tambahan Komunikasi Pemasaran Terpadu).* Jilid I, edisi Terjemahan, Jakarta: Erlangga

[10] Kasali, Rhenald, (2006) " Pop Marketing dlam Konteks Pemasaran Indonesia", Gramedia Pustaka Utama, Jakarta

[11] Nasrullah, R. (2016). Meme dan Islam : Simulakra Bahasa Agama di Media Sosial. *Academic Journal for Homiletic Studies* (*Vol. 10, Issue 1*), 113–128

[12] Feldman, Ronen, & Sanger, James. (2007). *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, New York

[13] Berry, M. W., & Kogan, J. (2010). Text Mining Application and Theory. United Kingdom: WILEY

[14] Taufik, I., & Pamungkas, S.A. (2018). Analisis Sentimen Terhadap Tokoh Publik Menggunakan Algoritma Support Vector Machine (SVM). Jurnal LOGIKA (Vol. 8, Issue 1). 69-79. https://doi.org/10.54367/means.v5i1.615

[15] Christianini, N., & Shawe-Taylor, J (2000) *Support Vectore Machines Cambridge,* Cambridge University Press, UK