SOCIAL MEDIA SENTIMENT ANALYSIS TO MEASURE COMMUNITY RESPONSE IN THE MILLENNIAL ROAD SAFETY FESTIVAL PROGRAM USING TF-IDF AND SUPPORT VECTOR MACHINE

Saiful Bukhori
Technology Information Department
Faculty of Computer Science
Universitas Jember
Jl. Kalimantan 37 Jember, East Java
saiful.ilkom@unej.ac.id

Sonya Sulistyono
Civil Engineering Department
Faculty of Engineering
Universitas Jember
Jl. Kalimantan 37 Jember, East Java
sonya.sulistyono@unej.ac.id

Antonius Cahya Prihandoko
Technology Information Department
Faculty of Computer Science
Universitas Jember
Jl. Kalimantan 37 Jember, East Java
antonius.cahya@unej.ac.id

Januar Adi Putra [1]
Technology Information Department
Faculty of Computer Science
Universitas Jember
Jl. Kalimantan 37 Jember, East Java
januaradi.putra@unej.ac.id

Windi Eka Yulia Retnani
Informatics Department
Faculty of Computer Science
Universitas Jember
Jl. Kalimantan 37 Jember, East Java
windi.ilkom@unej.ac.id

Umroh Makhmudah
System Information Department
Faculty of Computer Science
Universitas Jember
Jl. Kalimantan 37 Jember, East Java
makhmudahumroh@gmail.com

Muhammad Noor Dwi Eldianto
System Information Department
Faculty of Computer Science
Universitas Jember
Jl. Kalimantan 37 Jember, East Java
eldi.anto@gmail.com

ABSTRACT

This Sentiment Analysis is a combination of data mining and text mining. Sentiment Analysis itself is used to process various opinions that the public or experts have given through a variety of existing media. The argument is given to a product, service, or agency. Sentiment Analysis has three types of opinions: negative opinions, positive opinions, and neutral opinions. Based on the test results, the resulting model achieves the highest accuracy of 83.33% when using 80:20 scenario data, while the lowest accuracy of 80.00% is achieved when using the 60:40 scenario data. The higher the precision that will be obtained, whereas using less training data will be slightly unstable.

Keywords: Sentiment Analysis, SVM, TF, IDF, MRSF

INTRODUCTION

To increase driving safety awareness for young millennials, the Indonesian Police Traffic Corps collaborated with the National Movement Association (PGK). The Indonesian Traffic Volunteers (Relasi) held the Millennial Road Safety Festival Program on February 2-March 31, 2019. This activity was based on millennials in Indonesia, amounting to 65 million people from the age range of 20-35 years, which are recorded as the generation that accounts for the highest number of traffic accidents, reaching 55%.

[1] Coressponding author: J.A. Putra (januaradi.putra@unej.ac.id)

The Millennial Road Safety Festival program itself is held in a casual style to make the process of socialization and education about road safety more interesting. The 2019 Millennial Road Safety Festival Program is controlled with a roadshow to 34 provinces in Indonesia. This activity aims to decrease victims of traffic accidents in 2020.

The Millennial Road Safety Festival program certainly gets many responses from the public, especially on social media. The response can be in the form of positive, negative, or neutral responses. Positive opinions can be obtained because of community satisfaction with the program. In contrast, negative opinions can be caused by the lack or imperfection of information in these programs or the subjectivity factor. Responses related to the Millennial Road Safety Festival Program by the public on social media certainly need to be summarized and analyzed to find out. They can be the basis for decision-making for further programs.

One way to find out how people respond to a program is to summarize opinions on social media. Social media contains information, opinions, and input from the public about many things. Social media tends to be independent, and all netizens can express their opinions more freely. Social media that are widely used by the public include Facebook, Twitter, Instagram, and Youtube. The technique to summarize public responses on social media that can be done is Sentiment analysis.

Sentiment analysis or opinion mining refers to a broad field of natural language processing, linguistic computing, and text mining that aims to analyze the opinions, sentiments, evaluations, attitudes, judgments, and emotions of a person, whether the speaker or writer is concerned with a topic, product, service, organization, specific individuals, or activities. Sentiment Analysis is a combination of data mining and text mining. Sentiment Analysis itself is used to process a variety of opinions given by the public or experts through a variety of existing media. The opinion is given to a product, service, or agency. In Sentiment Analysis, there are three types of opinions: negative opinions, positive opinions, and neutral opinions. This research expects to get a summary of community responses on social media related to the Millennial Road Safety Festival Program, which can be a reference in making decisions in subsequent programs

METHOD

The research method in this study can be seen in Figure 1.

Social Media Comments

The data used in this study is a collection of data in the form of public comments on social media Twitter, Facebook, Youtube, Instagram, and Online News related to the Millennial Road Safety Festival Program.

Preprocessing

Preprocessing is done on social media comments first because not all the attributes in the comments column are used to analyze the problem. In preprocessing, there are several stages, namely:

1. Case Folding
   The process for converting all uppercase letters to comments into lowercase letters (Strauss, 1990).

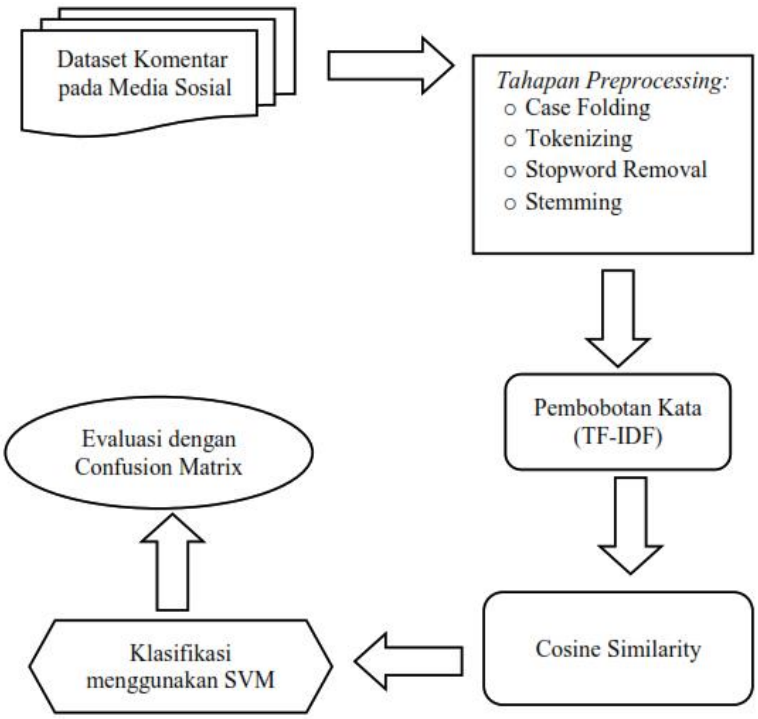

Figure 1. Research method

2. Tokenizing
   At this stage, the first character is checked to the last character. If the I character is not a word separator character such as dot (.), Comma (,), space, and other separators, then it will be combined with the next character (Kurniawan et al., 2014).
3. Stopword removal
   At this stage, each word will be checked for comments, then the process of eliminating words that are considered not important, such as there are conjunctions, prepositions, pronouns, or words that have nothing to do with Sentiment Analysis, then it will be deleted (Nurdiana et al., 2016).
4. Stemming
   This stage involves converting infix or suffix-filled words into a basic word that will contain more meaning to obtain information so that comments will be more specific in categorizing.

Word Weighting (TF-IDF)

The preprocessed dataset is then processed again into a binary number format so that the system can recognize it.

Cosine Similarity

Cosine similarity functions to compare the similarities between documents. In this case, what is compared is the query with the training document. In calculating cosine similarity, first, do a scalar multiplication between a query and a document, then add up, then do a multiplication between the length of the document and the length of the query that has been squared. After that, the square root is calculated, then the scalar multiplication results are divided by the results of the length of the document and query.

Classification Using Support Vector Machine

One statistical method that can be applied to classify is the Support Vector Machine (SVM). SVM is a technique for finding a hyperplane that can separate two data sets from two different classes (Vapnik, 1999).

Evaluate with the Confusion Matrix

After knowing the classification results using the Support Vector Machine, the next stage will be calculated accurately using the Confusion matrix, which is to calculate all test data that has been successfully classified according to the target. This will then do the accuracy calculation (comparing the cases identified correctly with the total number of cases).

System Design

The system design created includes Sentiment Analysis Process Architecture, Business Process Model Notation (BPMN), and Flowchart.

Architecture Sentiment Analysis Process

Sentiment analysis process architecture uses six processes: data collection process, manual labeling, preprocessing, feature vector, and sentiment classification, as illustrated in Figure 2 below.

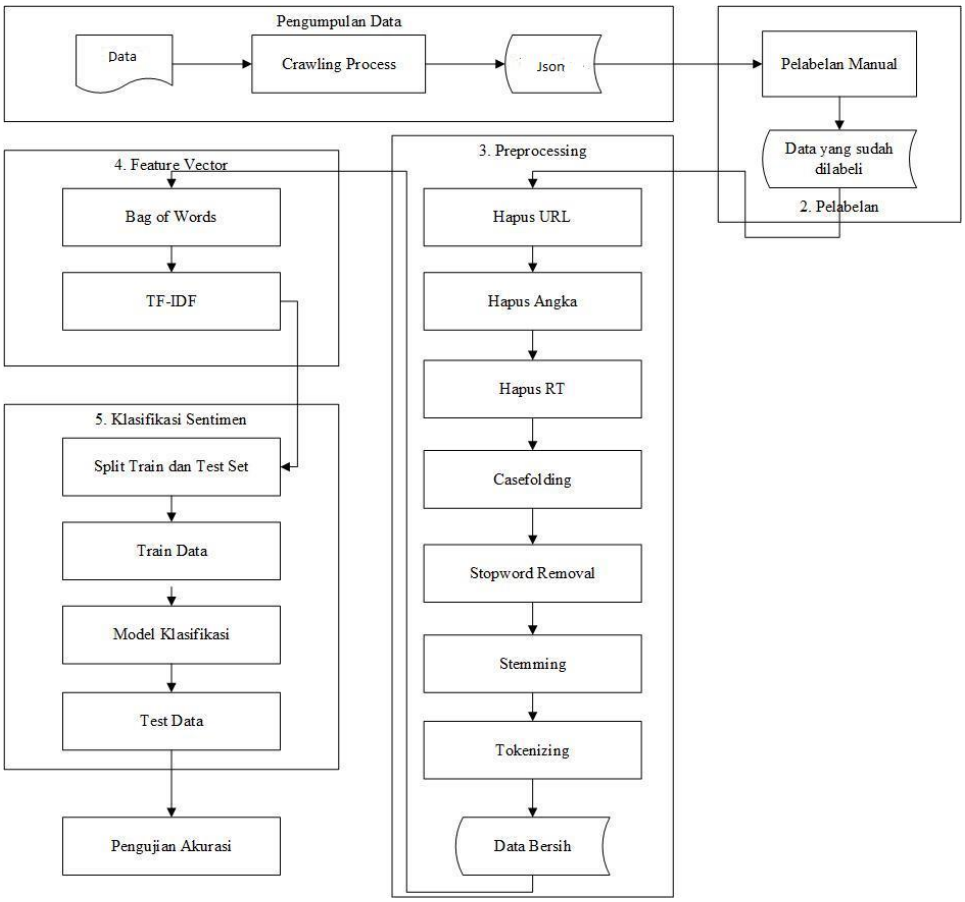

Figure 2. Sentiment analysis process

Business Process Model Notation (BPMN)

Business Process Model Notation (BPMN) is a graphical representation to determine the business processes of a business model; BPMN is made to make it easier for readers to know the business processes that occur. Here is the BPMN of this sentiment analysis system which can be seen in Figure 3 below.

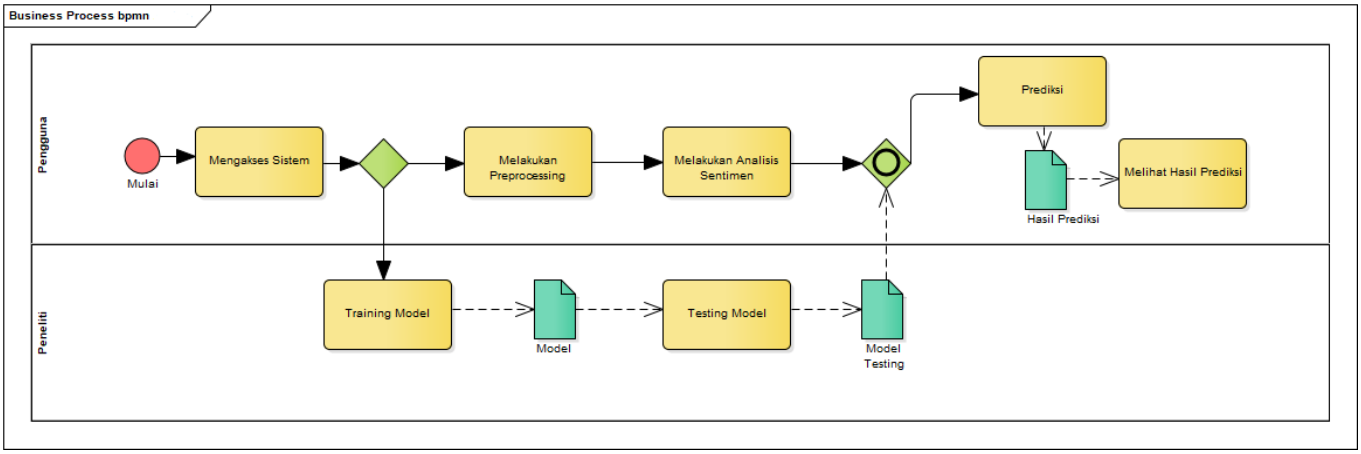

Figure 3. BPMN sentiment analysis

Flow chart

The flowchart is a chart consisting of various symbols and notations that are used to describe concisely and the algorithm of a system.

Crawling Flowchart

Flowchart Crawling describes the process of retrieving data from social media API sources. Data retrieval is performed on the data based on the desired query with the desired crawling time limit. Crawling data API uses authentication of consumer keys, consumer secrets, access tokens, and access secrets. The flowchart crawling process can be seen in Figure 4.

Flowchart Preprocessing

Flowchart preprocessing illustrates the flow of data processing before the sentiment analysis process using the SVM method. Flowchart Preprocessing can be seen in Figure 5.

Feature Extraction and Classification Process Flowchart

Flowchart classification process describes the Feature Extraction process, which is the process of obtaining features or changing text into the nominal form to be processed in the next process, the classification using the SVM method, the process flowchart can be seen in Figure 6.

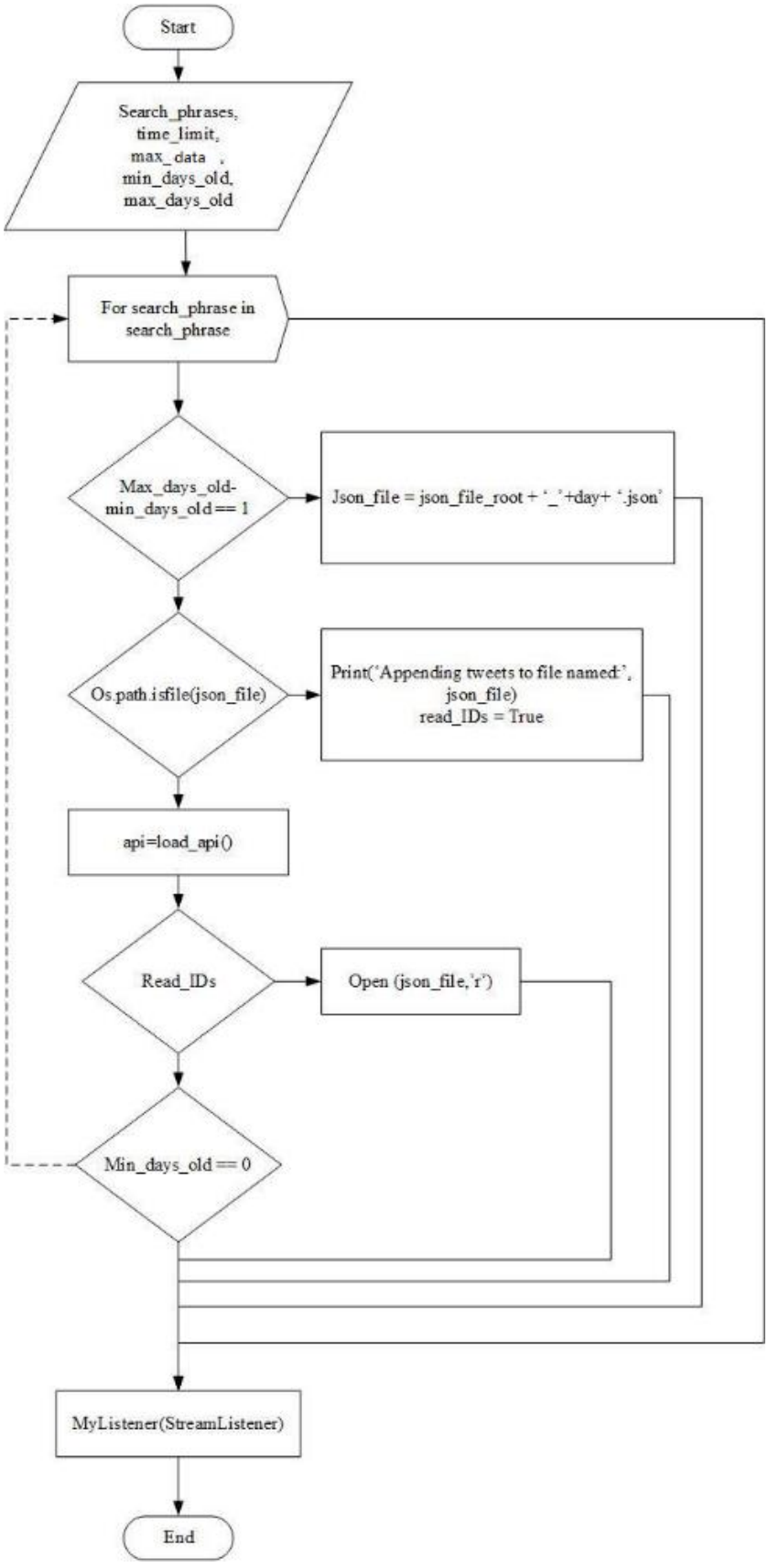

Figure 4. Crawling flowchart

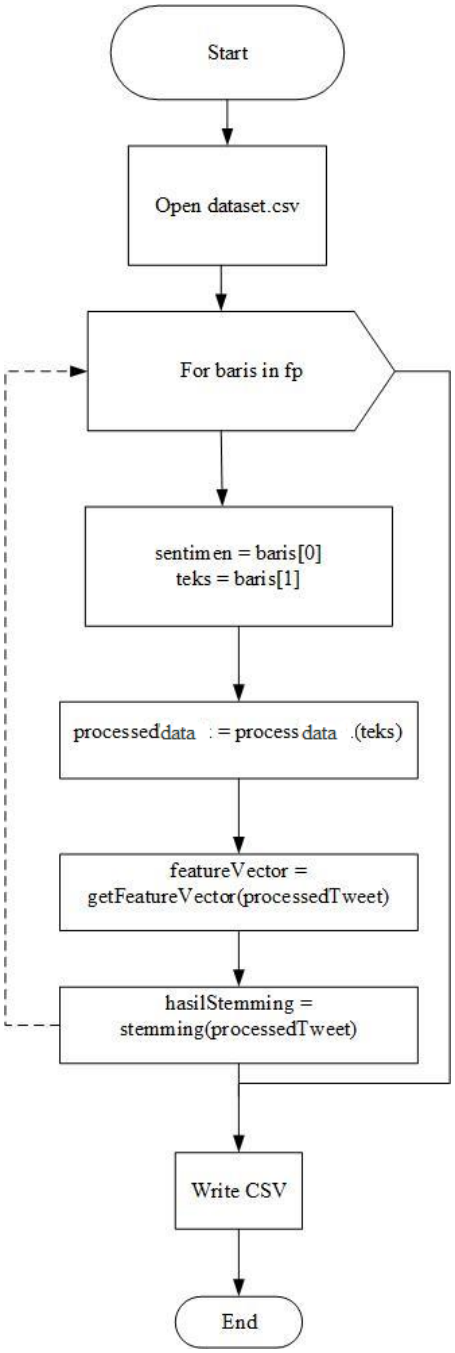

Figure 6. Preprocessing flowchart

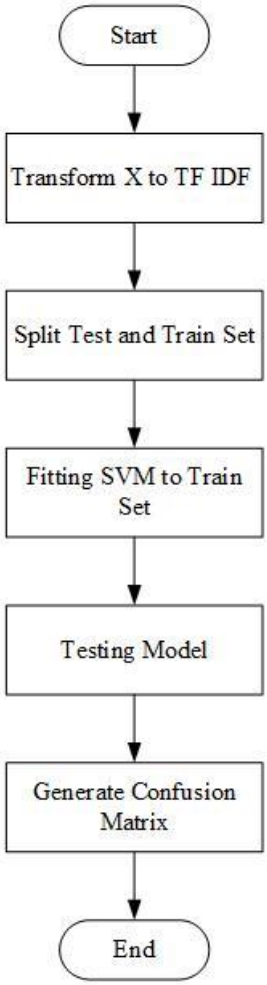

Figure 5. Classification flowchart

RESULT AND ANALYSIS

Research Dataset

The research dataset was taken from 5 sources, namely comments on online media, comments on Facebook social media, social media Twitter comments, Instagram social media comments, and YouTube social media comments. All comments are taken from posts

relating to Millennial Road Safety Festival activities. The maximum data retrieval limit is 30 October by crawling using the API of each social media while the online media is done manually. For social media, the author uses the hashtag #MRSF #MillenialRoadSafetyFestivel and #TemanLantas. The data classes used in this study are Neutral, Negative, and Positive, with the amount of data obtained is 352 Neutral, 202 Negative, and 311 Positive comments. In this study, the data used for each class only amounted to 200 comments. Online media sources used by the author and examples of comments used can be seen in Table 1 below.

Table 1. Dataset

No	Media	Link
1	Detik.com	https://news.detik.com/berita-jawa-timur/d4471168/millenial-road-safety-festival-di-suramadu-dinodailagu-jokowi-wae
2	Beritasatu.com	https://www.beritasatu.com/politik/543456/disusupikampanye-polri-pertimbangkan-setop-mrsf
3	Jawapos.com	https://www.jawapos.com/politik/18/03/2019/adakampanye-jokowi-di-acara-mrsf-2019-polri-itu-spontanwarga
4	Kumparan.com	https://kumparan.com/@kumparannews/ada-lagu-jokowiwae-saat-millennial-road-safety-festival-di-surabaya1552802251630168081
5	Infosurabaya.id	https://infosurabaya.id/2019/03/18/ada-kampanyeterselubung-di-millenial-road-safety-festival/
6	Radarjurnal.com	https://radarjurnal.com/2019/03/18/millenial-road-safetyfestival-di-suramadu-ternoda-aksi-kampanye-jokowi-wae/
7	Nusanews.id	https://www.nusanews.id/2019/03/millenial-road-safetyfestival-di.html
8	Mercusuar.com	https://berita-sosmed-indonesiapopuler.mercusuarumat.com/2019/03/millenial-road-safetyfestival-di.html
9	Beritasatu.com	https://www.beritasatu.com/politik/543453/kapolda-jatimtak-ada-niat-kampanye-dalam-acara-mrsf
10	Javanews.tv	https://www.javanews.tv/read/1566/kampanyekeselamatan-berkendara-suramadu-diwarnai-lagukampanye-politik
11	Aktual.com	https://www.aktual.com/polda-jatim-lepas-tangan-soalpeserta-mrsf-2019-yang-kampanyekan-jokowi/
12	Jawapos.com	https://www.jawapos.com/politik/18/03/2019/adakampanye-jokowi-di-acara-mrsf-2019-polri-itu-spontanwarga
13	Goriau.com	https://www.goriau.com/berita/baca/kampanye-jokowi-dimilenial-road-safty-festival-polisi-itu-spontan-warga.html
14	Gonews.co	https://www.gonews.co/berita/baca/2019/03/18/kampanye-jokowi-di-milenial-road-safty-festival-polisi-itu-spontanwarga
15	Beritasatu.com	https://www.beritasatu.com/politik/543677/mrsf-disusupikampanye-polri-tegaskan-netral-dalam-pemilu
16	Gosumbar.com	https://www.gosumbar.com/berita/baca/2019/03/18/kampanye-jokowi-di-milenial-road-safty-festival-polisi-ituspontan-warga

Table 2. Example comment dataset

No	Komentar Sosial Media	Class
1	Selamat pagi #TemanLantas jangan lupa hari Minggu 23 Juni 2019 pukul 06.00 WIB di Simpang Lima Semarang. "Merajut Persatuan dalam Kebhinekaan & Millennial Road Safety". Nikmati hiburannya dan Bawa pulang hadiahnya. . Gratissss! pic.twitter.com/k0NP5fjwv3	Neutral
2	Mohon maaf ini, banyak baliho besar program Milenial Road Safety pasang foto Jkw Bbrp hari kemarin lewat depan Polda Jabar, di barrier jalan terpasang baliho Millenial Road Safety dengan foto Jkw gede banget Kampanye ? Auk Ah Elap ~~~~	Negatif
3	Kira-kira... Andai yg diputar oknum peserta itu lagu #2019GantiPresiden apa masih akan dibilang 'itu ekspresi dia ?' #RinduPemimpinJujurDanAdil Millenial Road Safety Festival di Suramadu Dinodai Lagu 'Jokowi Wae'	Negatif
4	Keren nih acara Millenial Road Safety di Palembang. Di samping ada unsur edukasinya acaranya seru, rame dgn di hadiri lbh dr 100 rb warga juga menghibur	Positif
5	Anak2ku, generasi millennial NTB, setelah acara Millennial Road Safety Festival pagi tadi, tidak boleh lagi ugal2an di jalan raya, tidak boleh lagi ngebut di jalan raya. Ok	Positif

Preprocessing Implementation Results

At the text preprocessing stage, the dataset is taken by repeating it until the entire dataset is analyzed with four stages: case-folding, tokenization, filtering, and stemming. In the case-folding stage, the comment data that has been taken is converted into lowercase and deleting punctuation contained in the dataset and eliminating other components that do not affect the features and are not components in the analysis process. An example of the case-folding process can be seen in Table 3 below.

Table 3. Preprocessing result

No	Social Media Comments	Case-folding Result
1	Selamat pagi #TemanLantas jangan lupa hari Minggu 23 Juni 2019 pukul 06.00 WIB di Simpang Lima Semarang. "Merajut Persatuan dalam Kebhinekaan & Millennial Road Safety". Nikmati hiburannya dan Bawa pulang hadiahnya. . Gratissss! pic.twitter.com/k0NP5fjwv3	selamat pagi temanlantas jangan lupa hari minggu pukul wib di simpang lima semarang merajut persatuan dalam kebhinekaan millenial road safety nikmati hiburannya dan bawa pulang hadiahnya gratissss
2	Mohon maaf ini, banyak baliho besar program Milenial Road Safety pasang foto Jkw Bbrp hari kemarin lewat depan Polda Jabar, di barrier jalan terpasang baliho Millenial Road Safety dengan foto Jkw gede banget Kampanye ? Auk Ah Elap ~~~~	mohon maaf ini banyak baliho besar program milenial road safety pasang foto jkw bbrp hari kemarin lewat depan polda jabar di barrier jalan terpasang baliho millenial road safety dengan foto jkw gede banget kampanye auk ah elap
3	Kira-kira... Andai yg diputar oknum peserta itu lagu #2019GantiPresiden apa masih akan dibilang 'itu ekspresi dia ?' #RinduPemimpinJujurDanAdil Millenial Road Safety Festival di Suramadu Dinodai Lagu 'Jokowi Wae'	kira kira andai yang diputar oknum peserta itu lagu apa masih akan dibilang itu ekspresi dia millenial road safety festival di suramadu dinodai lagu jokowi wae
4	Keren nih acara Millenial Road Safety di Palembang. Di samping ada unsur edukasinya acaranya seru, rame dgn di hadiri lbh dr 100 rb warga juga menghibur	keren nih acara millenial road safety di palembang di samping ada unsur edukasinya acaranya seru rame dgn di hadiri lbh dr rb warga juga menghibur
5	Anak2ku, generasi millennial NTB, setelah acara Millen-nial Road Safety Festival pagi tadi, tidak boleh lagi ugal2-an di jalan raya, tidak boleh lagi ngebut di jalan raya. Ok	anakku gennerasi millenial ntb setelah acara millenial road safety festival pagi tadi tidak boleh lagi ugalan dijalan raya tidak boleh lagi ngebut di jalan raya ok

The dataset that has gone through the initial processing will go into the second processing, which is to stop stopwords or concluding words that do not contribute to the sentiment of a comment. The results of stopword removal or stopword removal can be seen in the following table 4:

Table 4. Stopword removal result

No	Social Media Comments	Stopword Removal Result
1	selamat pagi temanlantas jangan lupa hari minggu pukul wib di simpang lima semarang merajut persatuan dalam kebhinekaan millenial road safety nikmati hiburannya dan bawa pulang hadiahnya gratissss	selamat pagi temanlantas lupa simpang lima semarang merajut persatuan dalam kebhinekaan millenial road safety nikmati hiburannya bawa pulang hadiahnya gratiss
2	mohon maaf ini banyak baliho besar program milenial road safety pasang foto jkw bbrp hari kemarin lewat depan polda jabar di barrier jalan terpasang baliho millenial road safety dengan foto jkw gede banget kampanye auk ah elap	mohon maaf banyak baliho besar program milenial road safety pasang foto jkw kemarin lewat depan polda jabar barrier jalan teroasang baliho millenial road safety foto jkw gede banget kampanye auk elap
3	kira kira andai yang diputar oknum peserta itu lagu apa masih akan dibilang itu ekspresi dia millenial road safety festival di suramadu dinodai lagu jokowi wae	kira andai diputar oknum peserta itu lagu masih dibilang ekspresi dia millenial road safety festival suramadu dinodai lagu jokowi wae
4	keren nih acara millenial road safety di palembang di samping ada unsur edukasinya acaranya seru rame dgn di hadiri lbh dr rb warga juga menghibur	keren acara millenial road safety palembang samping unsur edukasinya acaranya seru rame hadiri lbh warga juga menghibur
5	anakku gennerasi millenial ntb setelah acara millenial road safety festival pagi tadi tidak boleh lagi ugalan dijalan raya tidak boleh lagi ngebut di jalan raya ok	anakku gennerasi millenial ntb setelah acara millenial road safety festival pagi tidak boleh ugalan dijalan tidak boleh ngebut jalan raya ok

The dataset that has gone through Stopword removal process will go into the third processing, which is to do the stemming process to make the whole words in each comment become the base words. The results of the stemming process can be seen in table 5 below:

Table 5. Stemming result

No	Social Media Comments	Stemming Result
1	selamat pagi temanlantas lupa simpang lima semarang merajut persatuan dalam kebhinekaan millenial road safety nikmati hiburannya bawa pulang hadiahnya gratiss	selamat pagi temanlantas lupa simpang lima semarang rajut satu bhineka millenial road safety nikmat hiburan bawa pulang hadiah gratiss
2	mohon maaf banyak baliho besar program milenial road safety pasang foto jkw kemarin lewat depan polda jabar barrier jalan terpasang baliho millenial road safety foto jkw gede banget kampanye auk elap	mohon maaf banyak baliho besar program milenial road safety pasang foto jkw kemarin lwat deoan polda jabar barrier pasang baliho millenial road safety foto jkw gede banget kampanye auk elap
3	kira andai diputar oknum peserta itu lagu masih dibilang ekspresi dia millenial road safety festival suramadu dinodai lagu jokowi wae	kira andai putar oknum peserta itu lagu masih bilang ekspresi dia millenial road safety festival suramadu noda lagu jokowo wae
4	keren acara millenial road safety palembang samping unsur edukasinya acaranya seru rame hadiri lbh warga juga menghibur	keren acara millenial road safety palembang samping unsur edukasi acara seru rame hadiri lbh warga menghibur
5	anakku gennerasi millenial ntb setelah acara millenial road safety festival pagi tidak boleh ugalan dijalan tidak boleh ngebut jalan raya ok	anakku gennerasi millenial ntb setelah acara millenial road safety festival pagi tidak boleh ugalan jalan tidak boleh ngebut jalan raya ok

TF-IDF Calculation

The TF IDF calculation process is used to provide weighting for the words to be sentiment analysis process. The calculation process will be carried out in the sample data in Table 6.

Table 6. TF-IDF Calculation

No	Text	Class
1	selamat pagi temanlantas lupa simpang lima semarang rajut satu bhineka millenial road safety nikmat hiburan bawa pulang hadiah gratiss	Neutral
2	mohon maaf banyak baliho besar program milenial road safety pasang foto jkw kemarin lwat deoan polda jabar barrier pasang baliho millenial road safety foto jkw gede banget kampanye auk elap	Negatif
3	kira andai putar oknum peserta itu lagu masih bilang ekspresi dia millenial road safety festival suramadu noda lagu jokowo wae	Negatif
4	keren acara millenial road safety palembang samping unsur edukasi acara seru rame hadiri lbh warga menghibur	Positif
5	anakku gennerasi millenial ntb setelah acara millenial road safety festival pagi tidak boleh ugalan jalan tidak boleh ngebut jalan raya ok	Positif

For example, if the data produces ten words for the TF IDF calculation process. Each word that has been formed weighs according to each document or sentence. The word gets one weight if it appears in a sentence and adds a D value from the word. After all the words have been weighted, it will go into the DF calculation process, which is the sum of the weight of a word in the whole document. DF will be the divisor for the total document analyzed. For this calculation, the total document used is five documents, so the D / DF is 5 divided by the value of DF. Simulation calculations can be seen in the following Table 7.

Table 7. Calculation simulation

Kata	TF					DF	D/DF
	D1	D2	D3	D4	D5		
Millenial	1			1		2	2.5
Road	1			1		2	2.5
Safety	1		1			2	2.5
Festival		1		1		2	2.5
Jokowi		1				1	5
Lagu			1			1	5
Suramadu			1		1	2	2.5
Meriah					1	1	5
Hiburan	1					1	5
Temanlantas		1			1	2	2.5

After getting the D / DF value, it goes to the next process to calculate the IDF value. The IDF value is obtained from the D / DF value log. After obtaining the IDF value, then the

value is added by one. The final process is calculating the value of W or the weight of each document or sentence. The value of W is obtained by multiplying the value of TF from each word for each document with the value IDF + 1. The total of W values for each document is W for the document. This calculation process can already prepare the dataset to be processed in the SVM classification algorithm. Examples of simulation calculations W can be seen in Table 8 below.

Table 8. Weight calculation

IDF	IDF+1	W=TF*(IDF+1)				
		D1	D2	D3	D4	D5
0.39794	1.397940009	1.39794	0	1.39794	1.39794	0
0.39794	1.397940009	1.39794	0	1.39794	1.39794	0
0.39794	1.397940009	1.39794	0	0	0	0
0.39794	1.397940009	0	1.39794	1.39794	1.39794	0
0.69897	1.698970004	0	1.69897	0	0	0
0.69897	1.698970004	0	0	0	0	0
0.39794	1.397940009	0	0	0	0	1.39794
0.69897	1.698970004	0	0	0	0	1.69897
0.69897	1.698970004	1.69897	0	0	0	0
0.39794	1.397940009	0	1.39794	0	0	1.39794
TF-IDF Result		5.89279	4.49485	4.19382	4.19382	4.49485

The results of the calculation of the weight generated by each word for each sentence become the basis of the calculation process as an X value for the classification process using SVM.

Accuracy Results with Data Scenarios

The testing process with new data is carried out in order to find out the actual accuracy of the model. In this process, testing is carried out using several data scenarios. Following are the results of testing in various models through several data scenarios.

Table 9. Accuracy result

Confusion Matrix		Predicted Class for 60 : 40 Accuracy			Predicted Class for 70 : 30 Accuracy			Predicted Class for 80 : 20 Accuracy		
		Negatif	Neutral	Positif	Negatif	Neutral	Positif	Negatif	Neutral	Positif
Actual Class	**Negatif**	64	12	4	46	10	4	30	8	2
	Neutral	0	69	11	0	52	8	0	35	5
	Positif	3	18	59	3	7	50	3	2	35

1. Testing results with 60:40 data scenario
   The table above shows the results of testing using a model that has been produced with a 40:60 data scenario showing that 64 data were accurately predicted for the negative class, 69 for the neutral class, and 59 data were accurately predicted for the positive

class. From the experiments conducted, 48 data were predicted to be inaccurate. The accuracy for this data scenario is 0.80 or 80.00%.

2. Testing results with 70:30 data scenarios
   The table above shows the results of testing using a model that has been produced with 30:70 data scenarios showing that 46 data are accurately predicted for negative classes, 52 for neutral classes, and 50 data are accurately predicted for positive classes. And 32 data were predicted inaccurately. The accuracy for this data scenario is 0.8222222 or 82.22%.
3. Testing results with 80:20 data scenarios
   The 80:20 data scenario has an accuracy of 0.833333 or 83.33%. Based on the table above, the test results using a model that has been produced with 80:20 data scenarios show that 30 data are accurately predicted for the negative class, 35 for the neutral class, and 35 data are accurately predicted for the positive class. And 20 data were predicted inaccurately.

System Interface

Figure 7. System interface software

CONCLUSION

1. This sentiment analysis system can classify social media comments with fairly good accuracy. This system can classify sentiments into three types, namely positive, neutral and negative.
2. The data classes used in this study are Neutral, Negative, and Positive, with the amount of data obtained by the end of October 2019 of 352 Neutral, 202 Negative, and 311 Positive comments.
3. Based on the test results, the resulting model achieved the highest accuracy of 83.33% when using 80:20 scenario data, while the lowest accuracy of 80.00% was achieved when using the 60:40 data scenario.
4. Data scenarios can affect the level of accuracy the more amount of training data provided, the higher the accuracy that will be obtained, whereas if you use less training data, the results will be a little unstable.
5. The *balanced* dataset in the study did not use the *oversampling* process to maximize the performance of the model.

ACKNOWLEDGEMENT

Thanks to the Directorate of Security and Safety of the National Police Traffic Corps for supporting research funding through the Traffic Accident Research Center.

REFERENCES

Kurniawan, A. Solihin, F., and Hastarita, F. 2014. Perancangan dan Pembuatan Aplikasi Pencarian Informasi Beasiswa dengan Menggunakan Cosine Similarity. Jurnal Simantec Volume 4,Nomor 2.

Nurdiana, O., Jumadi., dan Nursantika, D. 2016. Perbandingan Metode Cosine Similarity dengan Metode Jaccard Similarity pada Aplikasi Pencarian Terjemahan Al-Qur'an dalam Bahasa Indonesia.Jurnal Online Informatika Volume 1, Nomor 1.

Strauss, A. L. 1990. Basics of qualitative research. Grounded theory procedures and techniques. USA: Sage Publications.

Strauss, A. L. & Corbin, J 1997. Grounded theory in Practice. Thousand Oaks: Sage. Thousand Oaks: Sage.

Vapnik, V dan Cortes, C. 1995. Support Vector Networks. Machine Learning, 20, 273-297.