P-ISSN: 2685-8738 E-ISSN: 2685-872X

Variance: Journal of Statistics and Its Applications
Volume 5 Nomor 1 | April 2023
10.30598/variancevol5iss1page89-98

TEXT CLUSTERING ANALYSIS FOR PUBLIC SENTIMENT BASED ON TWITTER OPINIONS ON E-TILANG

F. Pattiasina[1*], W. A. Renwarin [2], N. F. Umar[3], Fadila[4], E. B. Ketyaru[5]
[1,2,3,4,5]Departement of Statistics, University of Pattimura, Ambon, 97233, Indonesia

Jl. M. J. Putuhena, Poka, 97233, Ambon-Maluku

E-mail Corresponding Author: *pattinamaf@gmail.com*

Abstract: Electronic Tickets (E-Tilang) have the advantage of faster service and have a very practical and fast system. Implementation of the E-Tilang system to facilitate transparency in the implementation of the ticket process or as a substitute for the ticket process on the spot. The effectiveness and efficiency of the E-Tilang system has generated various comments from the public. This study aims to determine the public's perception (sentiment) about E-Tilang. The data used in this study comes from people's tweets on Twitter about electronic tickets. This research shows that from the 524 Tweet data obtained, 252 people have positive sentiments, there are 98 people who have negative sentiments and there are people who have 174 neutral sentiments towards E-Tilang.

Keywords: Sentimen analysis, E-Tilang, Clustering.

1. INTRODUCTION

One of the innovative steps in renewal and change taken by the Indonesian government to improve the public service system for traffic violations is by implementing an electronic ticketing system [1]. This system is a solution to disciplining motorized vehicle drivers from committing many traffic violations [2]. In 2021 the National Police recorded that there were 2.12 million traffic violations in Indonesia. Based on data from the Indonesian Police's National Criminal Information Center, Among the 2.12 million traffic violations, there are 879,962 serious violations, 269,996 moderate violations and 965,286 minor violations [3].

The information system for every violation by motorists on the highway must be able to become the basis for prosecution for violations in the next stage, meaning that information on violations that have been committed by each person must always be identified by all members of the police who carry out the ticket. It's no secret that the practice of bribery during traffic operations often occurs [4][5]. So that is the reason that can be used as a basis for the National Police to start implementing a new system called the E-Tilang system. E-Tilang is a digitization of the ticketing process, by utilizing technology it is hoped that the entire ticketing process will be more efficient. In Article 272 of Law No. 22 of 2009 concerning Road Traffic and Road Transportation, it is stated that to support violation enforcement activities in the field of traffic and road transportation, electronic equipment can be used [6].

E-tickets have the advantage of being faster than conventional ticketing. The advantage is that this system is very practical and fast. The application of the E-Ticket system is to facilitate speed and convenience, transparency in the implementation of the ticket process or as a substitute for the ticket process on the spot. Some of the benefits for traffic violators with the E-Tilang system are transparency of the actions of public officials in government administration activities, community empowerment where the community is expected to be able to transmit an attitude of orderly traffic after knowing the existing regulations to those around them so as not to violate existing regulations [7]. The responsiveness of the authorities will be higher and more responsive and more responsive to public complaints in terms of traffic and justice where every violator who commits the same violation will receive the same fine or punishment without discrimination [8][9].

Submitted: 2023-09-17 *Accepted: 2023-09-19*

The effectiveness and efficiency of the E-Tilang system has generated various comments from the public. The rapid development of social media, more and more people are writing their opinion about something. Twitter is a social media for sharing ideas, ideas, expressions and giving opinions in the form of short texts, photos or videos called tweets about a product, service, political issue, or viral and other things [10][11].

This prompted this research to use Twitter social media as a source of data to be analyzed in the form of sentiment analysis which is a process of understanding, extracting and processing textual data automatically to obtain information sentiment contained in opinion sentences [12].

One way to do sentiment analysis is to use data from social media. Thousands of submissions occur daily on every social media. Everyone can express their opinion through social media freely. These opinions contain positive, negative and neutral sentiments on a topic. Positive sentiment expresses a good opinion about a context, negative sentiment expresses a bad opinion in a context, while neutral sentiment expresses things that are not good or bad [13][14].

Collecting data information from Twitter can be done by integrating Website Scraping and Orange Information Mining tools. To make it easier to recognize the type of content from some tweet information, a text mining process is required for the tweet data by applying clustering techniques [15]. To group textual data based on the similarity of content that is owned into several clusters, so that each cluster contains textual data with similar content by using clustering techniques.

2. RESEARCH METHOD

2.1 Method of collecting data

The method used in this study is the experimental method by observing the variables as the object under study. This experimental method is a study that evaluates a certain condition to be controlled so that one or several variables can be controlled. The purpose of this study is to determine negative sentiment, positive sentiment, and neutral.

2.2 Research procedure

This research was conducted in several stages, among others.

1. Retrieval of raw data in the form of tweets using Twitter's API. Before collecting data, we must create a Twitter API developer account.
2. Identify the details of the data that will be used later to prepare for machine learning needs.
3. After analyzing the characteristics of the data, design and implementation are carried out using features that are processed using a Twitter widget to collect data from Twitter according to the machine learning workflow based on that data.
4. At this stage, crawl data from the Orange Data Mining application through the API with queries that must be filled in according to the topic applied in this study, namely e-tickets.
5. Doing preprocessing of raw data in the form of .txt will be filtered into data in the form of quality data that is easy to understand and structured according to research needs.
6. Then we enter the World cloud stage, at this stage we can see the data display and can filter out words that are unclear or not important.
7. Next we enter the topic modeling stage, at this stage unclear words or data are marked with a color so that we can find out which words or data can be deleted.
8. After that we insert the Heat Map widget so that in this process we have 10 topics and from topic 1 to topic 10 we can see several people who have positive, negative, neutral arguments.
9. When we have reached the data filtering stage and deleted data that is not important or unclear, then we enter the sentiment analysis. Where at this stage we can choose the Liu Hu method, we try to use Liu Hu.

10. Then we click the Tweet Profiler icon. In this section, we can choose the available emoticons to make it easier for us to determine sentiment.
11. Then we connect the topic modeling with the Heat Map so that we can see clustering with diagram-like results.
12. Then we connect the topic modeling with the Heat Map so that we can see clustering with diagram-like results.
13. Then after that we place it in the Box Plot so that we can see comparisons of different sentiments. So that it can be proven by the tweets or tweet data that we get with various pros and cons and can achieve the results that we will conclude.
14. Then we connect the topic modeling with the Heat Map so that we can see clustering with diagram-like results.

3. RESULTS AND DISCUSSION

The use of Orange Data Mining displays the Text Clustering Widget Design which is presented in the process flow as shown in Figure 1.

a. Widget structure in orange application

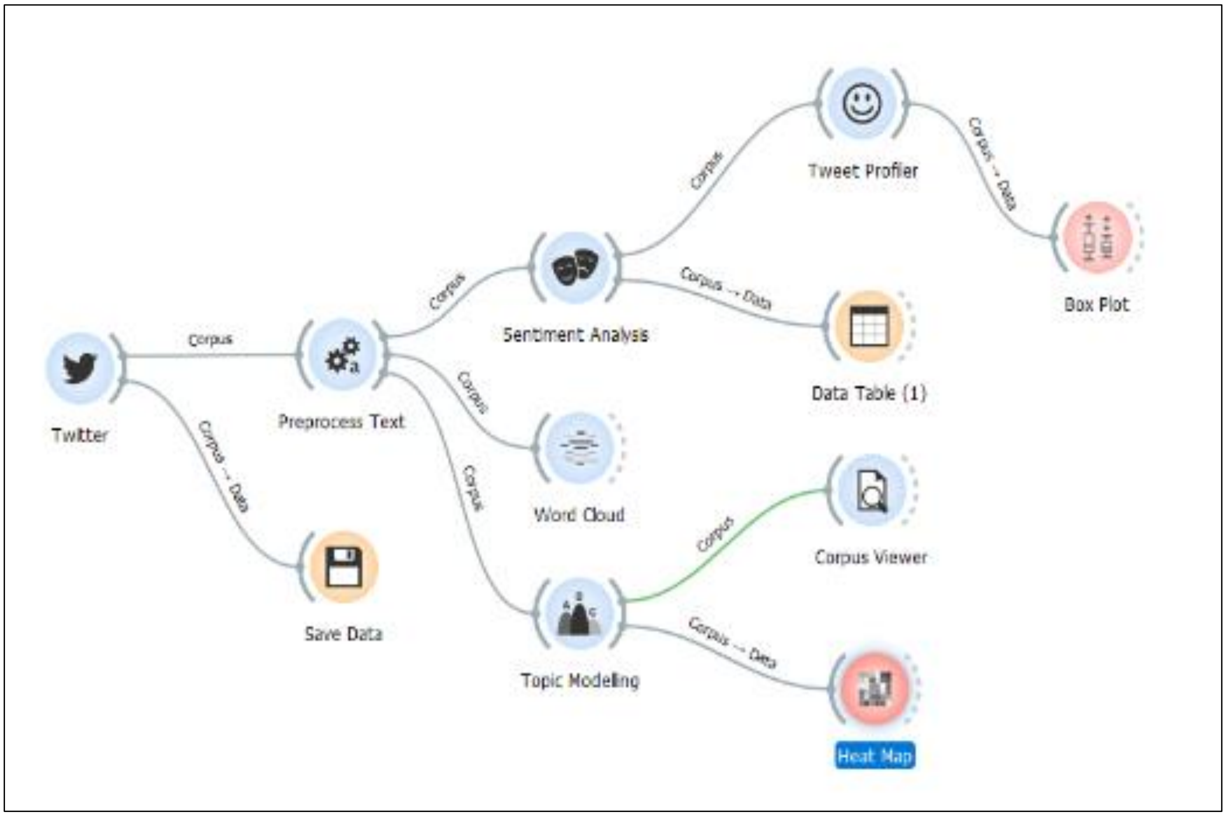
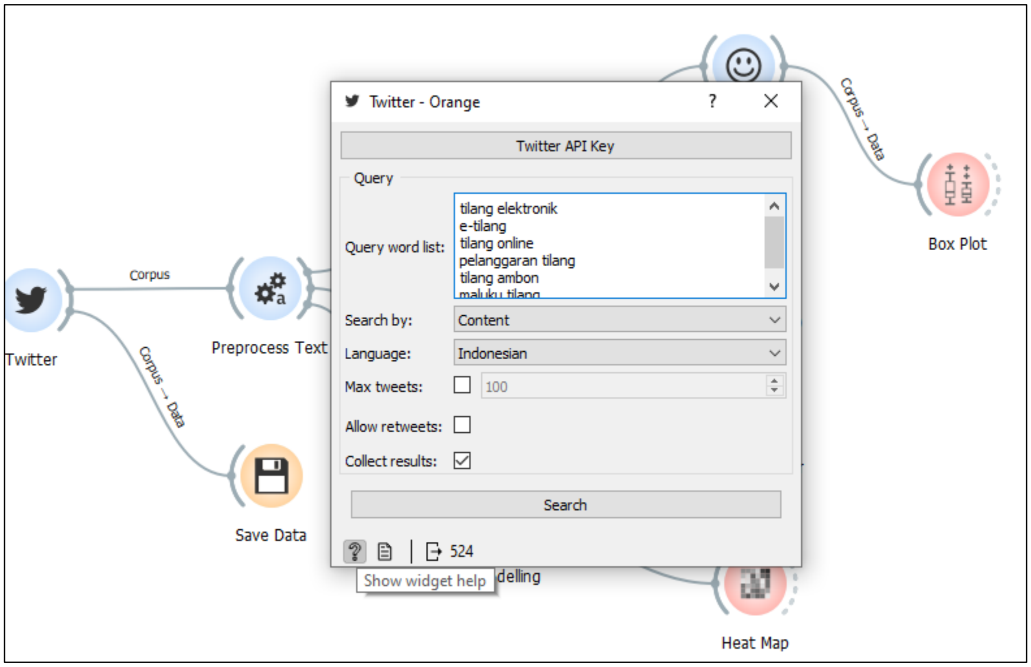


Figure 1. Design Widget Text Clustering

Data that has been crawled on Twitter will be inputted and analyzed one by one based on the object. Furthermore, it will be connected to the widgets needed for research so as to produce a widget design like the picture above.

b. Data Crawling (twitter)

In this research, the research data is public sentiment on Twitter regarding electronic ticketing, which is a recent government policy. A total of 524 tweets were obtained with the hashtags # and # electronic ticket #e-ticket #tilang online #pelanggarantilang #tilang ambon #maluku ticket from API and twitter key token. Where the results will be analyzed into clusters and recapitulated so as to create the dominant word arising from the status and opinions/comments. In Orange Data Mining the input data can be called a corpus. The corpus widget is a collection of documents that can present the number of lines of sentences, as well as determine which features will and will not be input for analysis purposes.



Twitter - Orange

Twitter API Key

Query

Query word list: tilang elektronik
e-tilang
tilang online
pelanggaran tilang
tilang ambon
maluku tilang

Search by: Content

Language: Indonesian

Max tweets: □ 100

Allow retweets: □

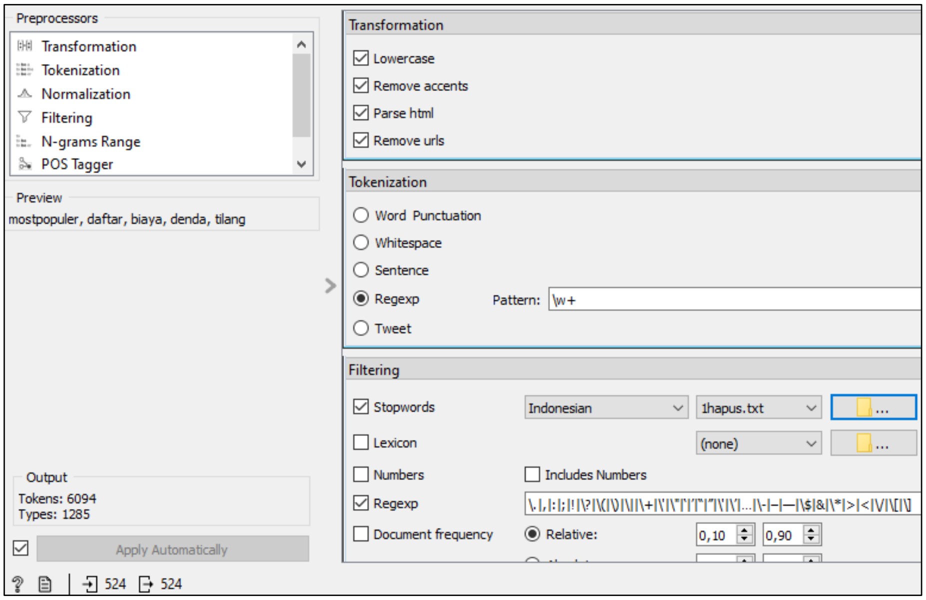
Collect results: ☑

Search

524

Figure 2. Data text mining (twitter)

c. Preproces text



Preprocessors

- Transformation
- Tokenization
- Normalization
- Filtering
- N-grams Range
- POS Tagger

Preview

mostpopuler, daftar, biaya, denda, tilang

Output

Tokens: 6094
Types: 1285

☑ Apply Automatically

524 524

Transformation

☑ Lowercase
☑ Remove accents
☑ Parse html
☑ Remove urls

Tokenization

○ Word Punctuation
○ Whitespace
○ Sentence
● Regexp Pattern: \w+
○ Tweet

Filtering

☑ Stopwords Indonesian 1hapus.txt
□ Lexicon (none)
□ Numbers □ Includes Numbers
☑ Regexp \.|,|:|;|!|\?|\(|\)|\||\+|\'|\"|'|'|"|"|\'|\'|…|\-|–|—|\$|&|*|>|<|\/|\[|\]
□ Document frequency ● Relative: 0,10 0,90

Figure 3. Preprocess Text

Before carrying out text analysis, the implementation of text mining to process text in this case is text preprocessing. Text is separated into smaller units (tokens), then transformation, tokenization, normalization, and filtering. Steps in the analysis applied sequentially can be enabled and disabled in Orange Data Mining in the Preprocess Text widget. Below are the stages in the preprocessing of the text before the text is analyzed using Orange Data Mining.

a. Transformation

The first thing to do from preprocessing text is transformation, which is the process of changing the input data to lowercase transformation by default.

1. Lowercase is used to convert all text to lowercase.

2. Remove Accents to remove all accents in the text.

3. Parse html to find out the html tags and parse the text.

4. Remove url to remove url from text.

b. Tokenization

Method of breaking text into smaller components (words, sentences, bigrams). a. Word & Punctuation will divide the text word by word and leave punctuation symbols as well. This example = (This), (example), (.) This process is done after the transformation.

c. Normalization

The next process is normalization which applies stemming and text lemmatization. Text that has been separated word for word will produce a text that stands alone in a sentence. Status content and opinions generally have imperfect arrangements (typos). It is hoped that with this normalization process the meaning of the text will be known by using WordNet Lemmatizer practicing a network of cognitive synonyms (synonyms) for tokens (words) based on the large Indonesian language lexicon (dictionary) information base from NLTK (Natural Language Toolkit).

d. Filtering

Delete or put word options. Here is a process in which the process of filtering words, symbols that are not needed in the next process (sentiment analysis).

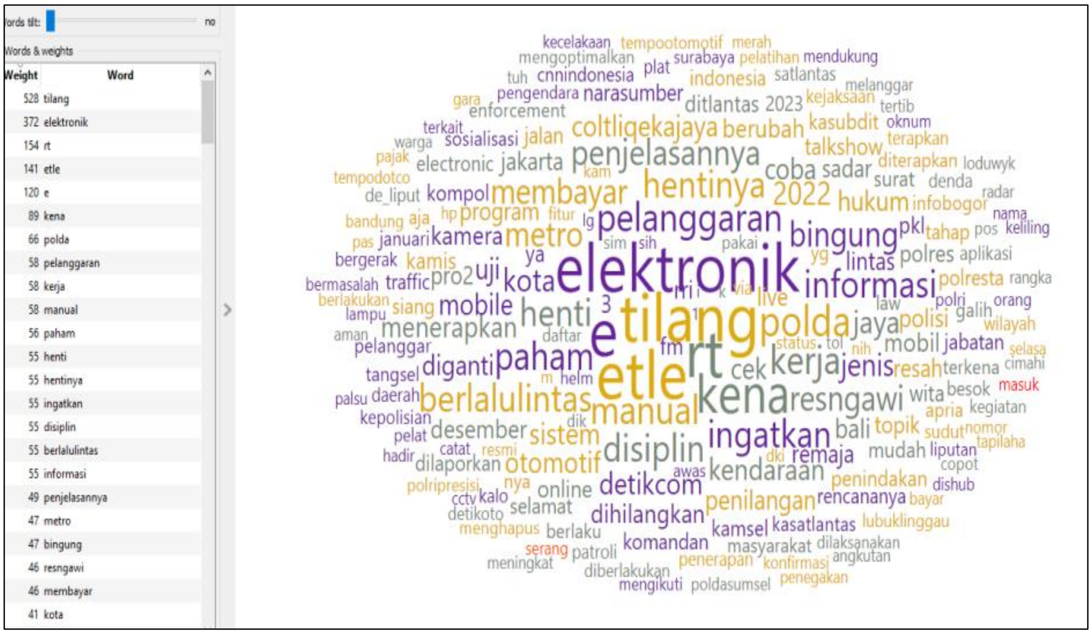
1) Stopwords Removes stopwords from reading (for example, removing and, or, this...). This can be tried by sorting out the language to be filtered (filter). The NLTK server provides downloadable stopwords for language purposes. However, in this research, default stopwords are in English. To filter the keywords provided.

2) Regexp removes words that match the regular expression. The default is set to remove punctuation.

3) Most Frequent Token, is text separated word by word. This feature will determine how many tokens will appear and will be analyzed in a document (cospus). In this research set 1, 000, 000 as the Most Frequent Token.

4) Process word cloud.

The text preprocessing stage has been carried out, then the data is in the form of separate texts and can be viewed in the form of a word cloud on Orange Data Mining. Word cloud can be used to find out the frequency of occurrence of the most words. Word cloud shares a display similar to the image above after going through the preprocess text. In the results of the image display above, the word cloud with the most frequency of words. If the frequency of word occurrences increases, the dimensions of the letters in the word cloud also become larger. Word cloud is an alteration or variation to show the results of the preprocess text stage. The style of the words makes the display more attractive and easy to understand. The image above is the result of the preprocess text stages of information – information which previously contained sentence lines for comments on e-ticket tweets throughout October 2022.



Words tilt: no

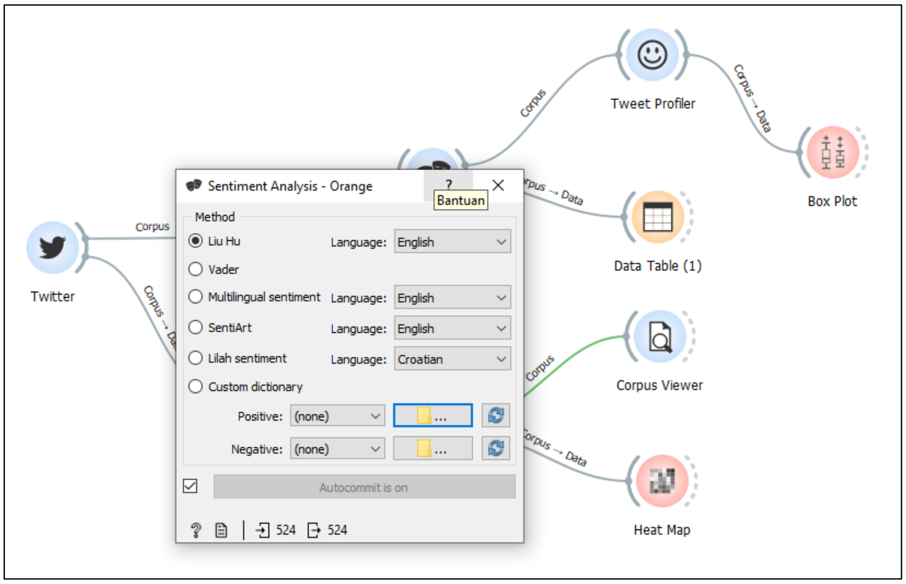
Words & weights

Weight	Word
528	tilang
372	elektronik
154	rt
141	etle
120	e
89	kena
66	polda
58	pelanggaran
58	kerja
58	manual
56	paham
55	henti
55	hentinya
55	ingatkan
55	disiplin
55	berlalulintas
55	informasi
49	penjelasannya
47	metro
47	bingung
46	resngawi
46	membayar
41	kota

kecelakaan tempootomotif merah
mengoptimalkan plat surabaya pelatihan mendukung
tuh cnnindonesia indonesia satlantas
gara pengendara narasumber melanggar
enforcement ditlantas 2023 kejaksaan tertib
terkait kasubdit oknum
warga sosialisasi jalan coltliqekajaya berubah talkshow terapkan
pajak electronic jakarta penjelasannya coba sadar diterapkan loduwyk
tempodotco kam surat denda
de_liput kompol membayar hentinya 2022 hukum infobogor radar
bandung aja hp program fitur lg pelanggaran nama
pas januari kamera metro sim sih bingung pkl tahap pos keliling
bergerak kamis ya pakai yg lintas polres aplikasi
bermasalah traffic pro2 uji kota elektronik informasi polresta rangka
berlakukan lampu siang mobile henti 3 tilang live law polri galih orang
aman menerapkan daftar polda jaya polisi wilayah
pelanggar e status tol nih mobil jabatan selasa
tangsel diganti paham fm rt cek kerja jenis resah terkena cimahi
palsu daerah m helm etle besok masuk
berlalulintas manual kena resngawi wita apria kegiatan
kepolisian dik bali topik sudut nomor
pelat desember sistem disiplin ingatkan tapilaha
hadir catat resmi dki remaja mudah liputan
dilaporkan otomotif awas kendaraan copot
polripresisi nya online detikcom penindakan dishub
cctv kalo penilangan rencananya bayar
detikoto selamat dihilangkan
menghapus berlaku kamsel kasatlantas lubuklinggau
serang patroli komandan masyarakat dilaksanakan
meningkat diberlakukan penerapan konfirmasi angkutan
mengikuti poldasumsel penegakan

Figure 4. Word Cloud

a. Sentiment analysis



Sentiment Analysis - Orange

Method

● Liu Hu Language: English

○ Vader

○ Multilingual sentiment Language: English

○ SentiArt Language: English

○ Lilah sentiment Language: Croatian

○ Custom dictionary

Positive: (none) ...

Negative: (none) ...

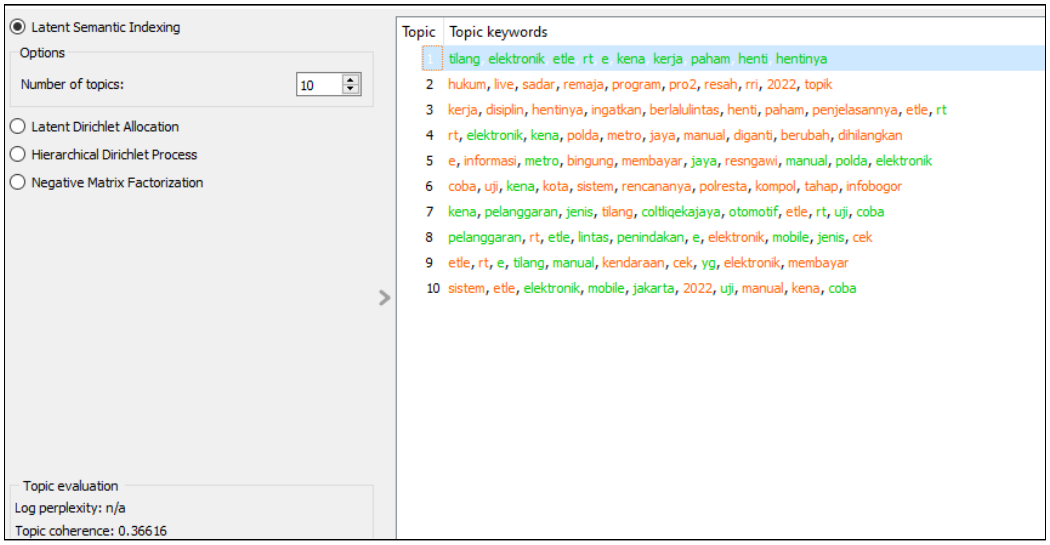
☑ Autocommit is on

524 524

Figure 5. Widget Sentiment Analysis Orange Data Mining

The analysis process uses the Vader algorithm to classify polarity (class sentiment) as positive, negative and neutral with a total score of compound. Vader will categorize, and give a text score based on the value of each word listed in Vader's lexicon. The final result of the evaluation is the total score, which is the compound. It is this total score that will be recaptured and the results compared. In analyzing there are several variable attributes that you want to focus on which of the features that are used to be analyzed (used features) of the corpus widget are the Text trans columns and the translated text that is in each of the comments, the aim of which is to get results in the form of positive attributes negative, neutral, and total score (compound).

a. Topic Modelling



● Latent Semantic Indexing

Options

Number of topics: 10

○ Latent Dirichlet Allocation

○ Hierarchical Dirichlet Process

○ Negative Matrix Factorization

Topic	Topic keywords
1	tilang, elektronik, etle, rt, e, kena, kerja, paham, henti, hentinya
2	hukum, live, sadar, remaja, program, pro2, resah, rri, 2022, topik
3	kerja, disiplin, hentinya, ingatkan, berlalulintas, henti, paham, penjelasannya, etle, rt
4	rt, elektronik, kena, polda, metro, jaya, manual, diganti, berubah, dihilangkan
5	e, informasi, metro, bingung, membayar, jaya, resngawi, manual, polda, elektronik
6	coba, uji, kena, kota, sistem, rencananya, polresta, kompol, tahap, infobogor
7	kena, pelanggaran, jenis, tilang, coltliqekajaya, otomotif, etle, rt, uji, coba
8	pelanggaran, rt, etle, lintas, penindakan, e, elektronik, mobile, jenis, cek
9	etle, rt, e, tilang, manual, kendaraan, cek, yg, elektronik, membayar
10	sistem, etle, elektronik, mobile, jakarta, 2022, uji, manual, kena, coba

Topic evaluation

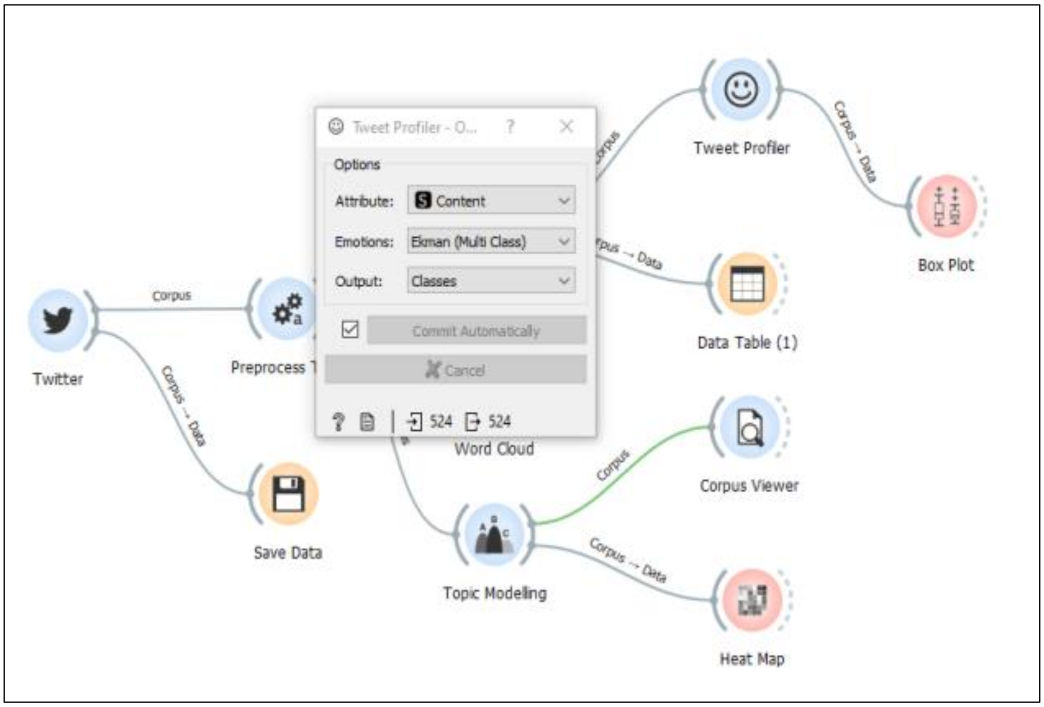
Log perplexity: n/a

Topic coherence: 0.36616

Figure 6. Topic Modelling

Topic modeling is a widget that can help us see the words or tweets that you want to delete. This stage is a widget for filtering words that are not accepted for analysis.

b. Tweet Profiler



Tweet Profiler - O...

Options

Attribute: Content

Emotions: Ekman (Multi Class)

Output: Classes

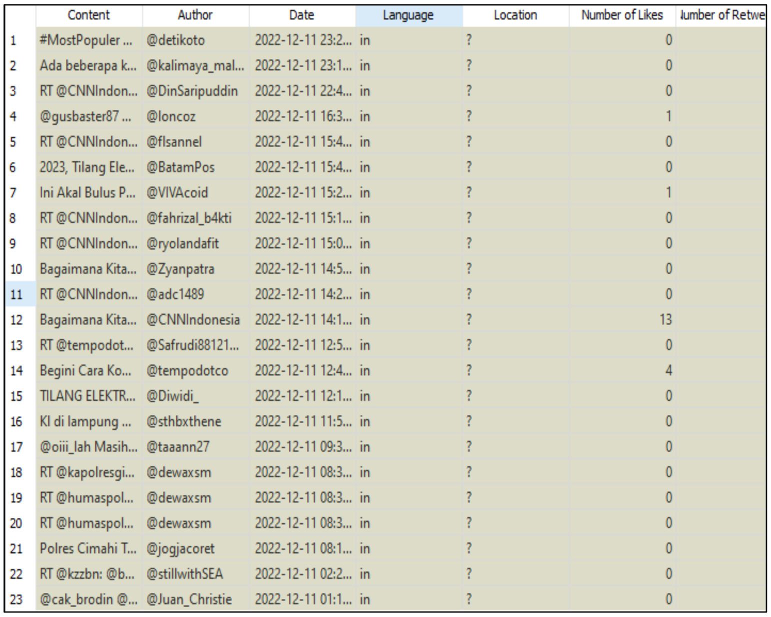
☑ Commit Automatically

Figure 7. Tweet Profiler

The Tweet Profiler fetches sentiment data from the server for every given tweet (or document). The widget sends information to the server, where the model calculates the probability and/or emotion score. The widget supports 3 emotion classifications, namely Ekmans, Plutchiks and Profile of Mood States (POMS). There are 3 categories of emotional classification, namely Ekmans, Plutchiks or Profile of Mood States. Multiple class classifications will result in one possible emotion per document, whereas multiple labels will create values in the column for each emotion. This research will use Content attributes for analysis, Ekman's classification of emotions with multi-class options and choose to observe the Emotions variables that have been grouped with orange data mining. In this study using data from 524 tweets regarding e-tilang. The data that has been crawled using the widget from orange data mining with Corpus and linked to the Tweet Profiler.

c. Table Data

The data table is also a result reaction to display which attribute is selected to be displayed as output in the select column widget on. previous steps. With the help of Twitter crawling data using the API analysis results in the sentiment analysis data table, it will be calculated how positive, negative, and neutral by looking at the total score (compound) with the calculation formula for numerical data formats.



	Content	Author	Date	Language	Location	Number of Likes	lumber of Retwe
1	#MostPopuler ...	@detikoto	2022-12-11 23:2...	in	?	0	
2	Ada beberapa k...	@kalimaya_mal...	2022-12-11 23:1...	in	?	0	
3	RT @CNNIndon...	@DinSaripuddin	2022-12-11 22:4...	in	?	0	
4	@gusbaster87 ...	@loncoz	2022-12-11 16:3...	in	?	1	
5	RT @CNNIndon...	@flsannel	2022-12-11 15:4...	in	?	0	
6	2023, Tilang Ele...	@BatamPos	2022-12-11 15:4...	in	?	0	
7	Ini Akal Bulus P...	@VIVAcoid	2022-12-11 15:2...	in	?	1	
8	RT @CNNIndon...	@fahrizal_b4kti	2022-12-11 15:1...	in	?	0	
9	RT @CNNIndon...	@ryolandafit	2022-12-11 15:0...	in	?	0	
10	Bagaimana Kita...	@Zyanpatra	2022-12-11 14:5...	in	?	0	
11	RT @CNNIndon...	@adc1489	2022-12-11 14:2...	in	?	0	
12	Bagaimana Kita...	@CNNIndonesia	2022-12-11 14:1...	in	?	13	
13	RT @tempodot...	@Safrudi88121...	2022-12-11 12:5...	in	?	0	
14	Begini Cara Ko...	@tempodotco	2022-12-11 12:4...	in	?	4	
15	TILANG ELEKTR...	@Diwidi_	2022-12-11 12:1...	in	?	0	
16	KI di lampung ...	@sthbxthene	2022-12-11 11:5...	in	?	0	
17	@oiii_lah Masih...	@taaann27	2022-12-11 09:3...	in	?	0	
18	RT @kapolresgi...	@dewaxsm	2022-12-11 08:3...	in	?	0	
19	RT @humaspol...	@dewaxsm	2022-12-11 08:3...	in	?	0	
20	RT @humaspol...	@dewaxsm	2022-12-11 08:3...	in	?	0	
21	Polres Cimahi T...	@jogjacoret	2022-12-11 08:1...	in	?	0	
22	RT @kzzbn: @b...	@stillwithSEA	2022-12-11 02:2...	in	?	0	
23	@cak_brodin @...	@Juan_Christie	2022-12-11 01:1...	in	?	0	

Figure 8. Table Data

d. Heat Map

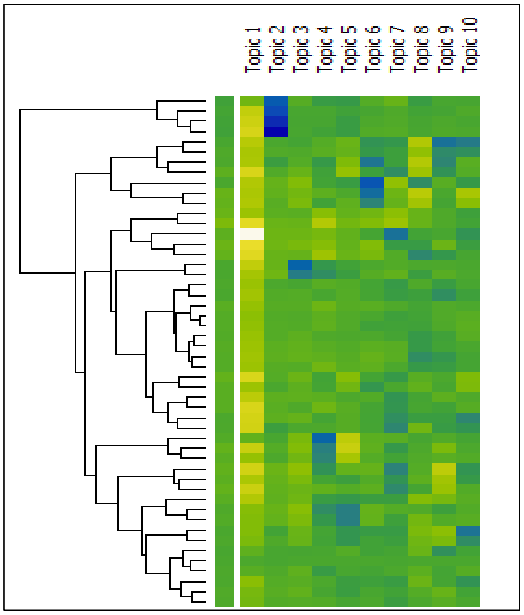


Figure 9. Heat Map

Heatmap is a visualization or mapping by displaying data with different color representations. Usually, the higher the number of a data group, the darker the color, generally symbolized by red. although it can be useful in various fields such as statistics, geography, and others.

e. Histogram

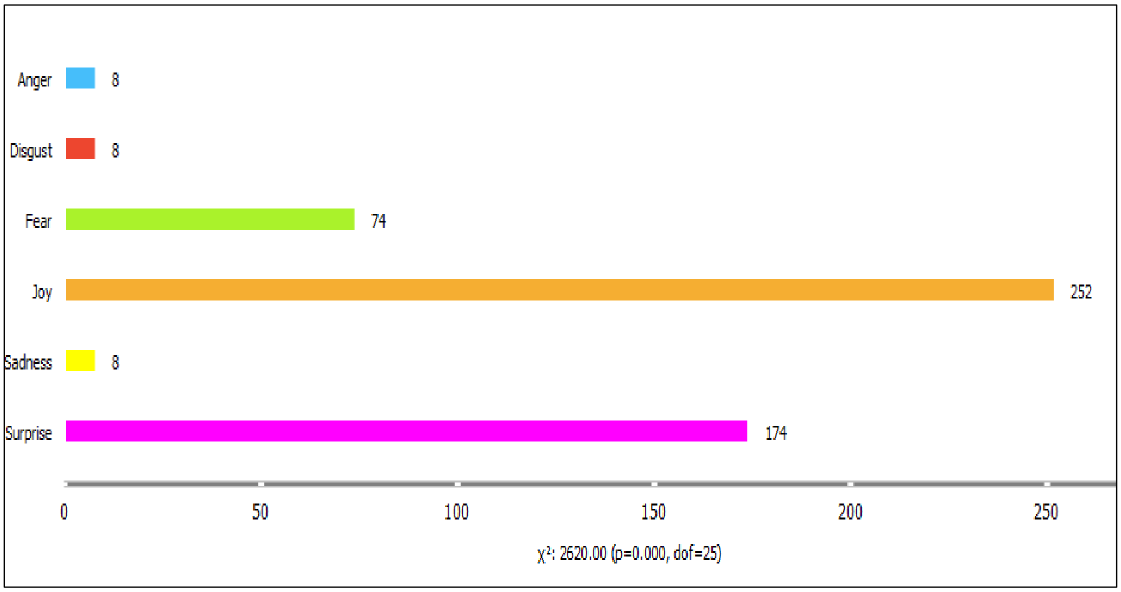


Figure 10. Histogram

The Box Plot widget shows the distribution of attribute values. It's good practice to check any new data with this widget to quickly find anomalies, such as duplicate values (eg gray or grey), outliers, and the like. So that according to the picture above, we can already conclude or analyze public sentiment regarding electronic fines. According to the existing data, there are several categories of sentiment.

a. Anger means that people are angry about electronic fines according to the results of the analysis of 8 people.

b. Disgust means that there are 8 people who don't like electronic ticketing related to the results of the analysis.

c. Fear means that people are afraid of electronic ticketing according to the results of an analysis of 74 people.

d. Joy means that people are happy about electronic fines according to the results of the analysis of 252 people.

e. Sadness means that people are sad about electronic fines according to the results of the analysis of 8 people.

f. Surprise means that the public is shocked by electronic ticketing according to the results of the analysis of 174 people.

.

4. CONCLUSIONS

Based on the results and discussion, it can be interpreted that of the 524 Tweet data obtained, it can be classified into 3 groups, namely Positive, Negative and Neutral sentiments with different opinions and sentiments. There were 252 people who had positive sentiments, 98 people had negative sentiments and there were 174 people who had neutral sentiments towards electronic tickets (E-Tilang).

REFERENCES

[1] Wulandari, A. (2020). Innovation In Implementing The E-Tilang System In Indonesia. Al-Mabsut, 1-10.

[2] Khalida, R., & Setiawati, S. (2020). E-Tilang System Sentiment Analysis Using Naive Bayes Algorithm With Information Gain Optimization. Journal of Information and Information Security (JIFORTY), 19-26.

[3] Sadya, S. (2022, December 12). Police Record 2.12 Million Traffic Violations in 2021. Retrieved from dataindonesia.Id: https://dataindonesia.id/ragam/detail/polri-catat-212-juta-pelanggaran-lalu- Lintas-pada-2021

[4] Setiyanto, Gunarto, & Wahyuningsih, S. (2017). The Effectiveness of Implementing E-Ticket Fines for Traffic Offenders Based on Law Number 22 of 2009 concerning Road Traffic and Transportation (Study at Rembang Police). Khaira Ummah Law Journal, 742-766.

[5] Matdoan, M. Y., & Igo, L. (2023). Application of x-means alghorithm for district/city clustering based on povetry rate in Maluku Islands and Papua. Jurnal Matematika dan Ilmu Pengetahuan Alam Lldikti Wilayah 1 (JUMPA), 3(1), 14-20.

[6] Rakhmadani, S. (2017). Analysis of The Application of E-Tilang in Realizing Good Governance in Indonesia. Proceedings of Social, Economics, and Humanities SNaPP 2017, 663-671.

[7] Ls, D., Lesnussa, Y. A., Talakua, M. W., & Matdoan, M. Y. (2021). Analisis Klaster untuk Pengelompokkan Kabupaten/Kota di Provinsi Maluku Berdasarkan Indikator Pendidikan dengan Menggunakan Metode Ward. Jurnal Statistika dan Aplikasinya, 5(1), 51-60.

[8] Troussas, C., Virvou, M., Espinosa, K. J., Llaguno, K., & Caro, J. (2013). Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning. Institute of Electrical and Electronics Engineers.

[9] Matdoan, M. Y., & Balami, A. M. (2019). Estimasi Parameter Regresi Kuantil Dengan Fungsi Spline Truncated Pada Kasus Demam Berdarah Dengue di Kota Surabaya. Jurnal Matematika dan Statistika serta Aplikasinya, 7(1), 44-53.

[10] Zuliana, V., Garno, & Maulana, I. (2022). Sentiment Analysis of Digital TV Migration Program Using Naive Bayes Algorithm with Chi Square. Journal of Information and Computers, 90-95.

[11] Lopies, C., Matdoan, M. Y., Loklomin, S. B., & Wattimena, A. Z. (2023). Analisis dan Klasifikasi Tingkat Kebahagiaan Masyarakat Berdasarkan Propinsi di Indonesia Dengan Pendekatan Statistik. PARAMETER: Jurnal Matematika, Statistika Dan Terapannya, 2(01), 157-169.

[12] Fitri, V., Andreswari, R., & Hasibuan, M. (2019). Sentiment Analysis of Social Media Twitter with the Case of Anti-LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm. Procedia Computer Science, 765–772.

[13] Colleta, L. F., da Silva, N. F., Hruschka, E. R., & Hruschka, E. R. (2014). Combining Classification and Clustering for Tweet Sentiment Analysis. Institute of Electrical and Electronics Engineers.

[14] Nasukawa, T., & Yi, J. (2003). Sentiment Analysis: Capturing Favorability Using Natural Language Processing. ACM Digital Libraries.

[15] Wiguna, R. A., & Rifai, A. I. (2021). Community Text Clustering Analysis on Twitter Regarding the Omnibus Law Using Orange Data Mining. Journal of Information Systems and Informatics.