JOURNAL OF MULTIDISCIPLINARY ISSUES
Journal Website: www.jmis.site
J. Multidisc. Issues 2(2) 1 - 21 (2022)

SENTIMENT ANALYSIS OF COMMENTS ON SEXUAL HARASSMENT IN COLLEGES ON FOUR POPULAR SOCIAL MEDIA

Vinson Phoan [a], Johan Setiawan [b]
[a] vinson.phoan@student.umn.ac.id, [b] johan@umn.ac.id
[a b] Universitas Multimedia Nusantara, Tangerang, Indonesia

ABSTRACT *Sexual harassment is a sexual act in the form of verbal and nonverbal that is carried out intentionally, and there are indications of coercion on the victim. Often the victim refuses to report or tell stories because the victim is afraid that something untoward will happen in the future. To find helpful information for survivors, in this study, we will get the sentiment on comments related to matters related to UMN by using the CRISP-DM framework method, FastText, and the SVM algorithm to identify a statement on sexual relations with UMN.*
This study used a model with an accuracy of 55.14% and sexual harassment data collected on 16 March 2022, with 287 data obtained from Twitter, Instagram, Medium, and Line Today sites. Positive sentiment is the least found sentiment from the overall data, only 8.7%, while negative sentiment is 36.6%. Of the four platforms, the best is the Twitter platform because it gets a pretty good response in terms of positive and neutral sentiments compared to others.
Objective – There are several objectives in this study, namely measuring the sentiment or response of netizens from each social media to cases of sexual harassment around the UMN environment, measuring the performance or performance of the implementation of the SVM algorithm on the topic of sentiment analysis from various social media on cases of sexual harassment-related to UMN, Find out how effective the hashtag #timestalk is on Twitter and Instagram.
Methodology – The method used in this research is to use the CRISP-DM framework, FastText, and SVM as a solution to the problem.
Findings – Using the SVM algorithm with a high accuracy level of 55.14%, implemented in the sexual harassment dataset related to UMN, found a genuine neutral sentiment of 54.7% or 157 comments, 36.6% or 105 negative sentiments, and 8.7% or 25 positive sentiments. Based on the SVM algorithm model, it was found that Twitter, Instagram, and Medium platforms get pretty good support when viewed from the frequency of words that appear primarily from neutral and positive sentiments.
Novelty – The difference between this research and the previous one is the implementation in specific and different cases, namely sexual harassment that occurred at UMN and using FastText to do word embedding.
Keywords: *Sexual Harassment; Sentiment Analysis; Natural Language Processing; Text Mining; Social Media*

JEL Classification: I23, K14, O35
Article Info: *Received 15 July 2022, Revised 06 August 2022, Accepted 07 August 2022*
Article Correspondence: vinson.phoan@student.umn.ac.id
Recommended Citation: Phoan V. & Setiawan J. (2022). Sentiment Analysis of Comments on Sexual Harassment in Colleges On Four Popular Social Media. Journal of Multidisciplinary Issues, Issues 2(2) 1-20

I. INTRODUCTION

Sexual harassment is a sexual act carried out intentionally and indicates coercion against victims who refuse (Myrtati D. Artaria, 2012).

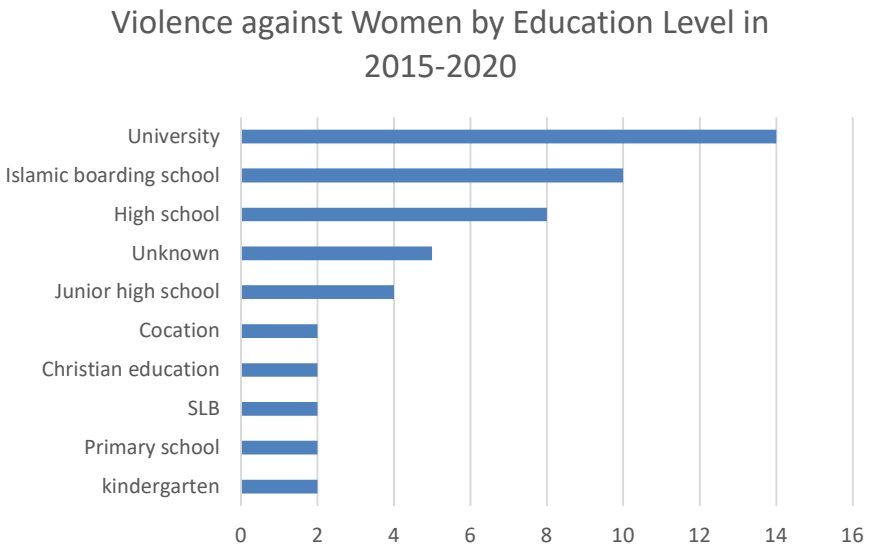


Figure 1 Graph of Violence against Women by Education Level in 2015-2020

Figure 1 is a graph of sexual violence and discrimination against women by education level 2015-2020 in Indonesia. The education level of the university ranked first with a total of 14 cases or 27%, followed by Islamic boarding schools occupying the second largest with 10 points or 19% (Komisi Nasional Perempuan, 2020).

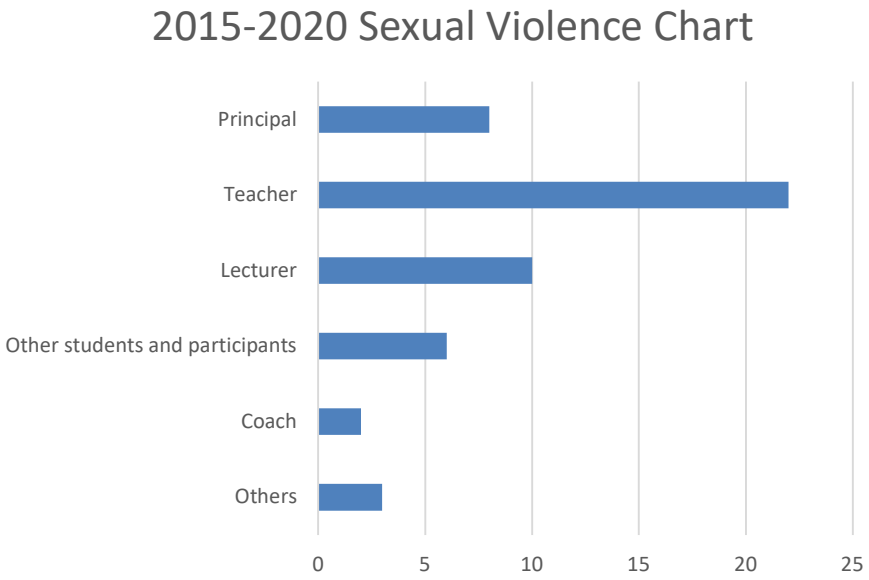


Figure 2 2015-2020 Sexual Violence Chart

Most forms of sexual violence from 2015 to 2020 were carried out in sexual form, with as many as 45 cases or 88% of 51 points, and it was known that the modus operandi was carried out by the perpetrators by inviting the victim to go out of town using excuses to do thesis research or students. In this case, the perpetrators of sexual violence were mainly carried out by teachers/ustadz with 22 patients. The lecturers

found ten issues out of 51 points and were the second-highest after teachers/ustadz (Komisi Nasional Perempuan, 2020).

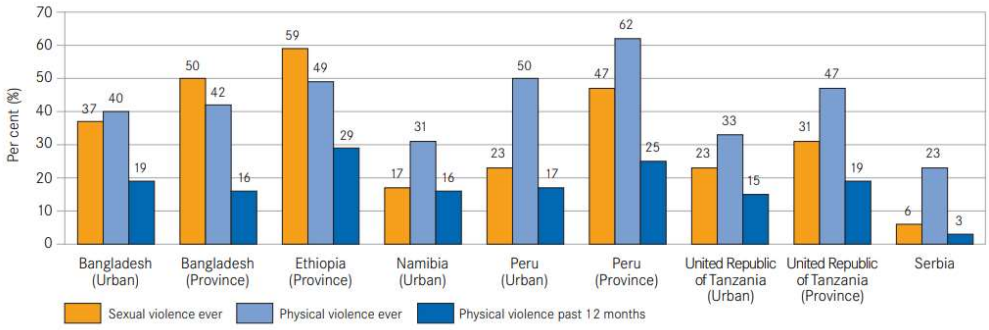


Figure 3 Violence in Relationships Chart

The same is the case with other countries (Bangladesh, Ethiopia, Namibia, Peru, Tanzania, and Serbia), based on previous research showing 18 SENTIMENT ANALYSIS OF SEXUAL HARASSMENT COMMENTS IN SOCIAL MEDIA: CASE STUDY UMN, Universitas Multimedia Nusantara there are still many perpetrators of sexual crimes in relationships which is where 23-56% of victims have experienced something similar, and there are several reasons why the victim does not leave or report to the perpetrator because the victim is afraid of revenge, economic support, no concern from the people around, and many other factors (World Health Organization, 2012). In addition, from the survey data, information about gender is mostly (90%) male perpetrators who are professors or lecturers themselves (Karami, White, et al., 2020).

This is evidenced by previous studies based on reports obtained from victims through the Everydaysexism website. From the prediction results, the shooting often occurs in the University environment or at home. This needs to be considered by the Minister of Education and Culture and the responsible parties (Karami, Swan, et al., 2020).

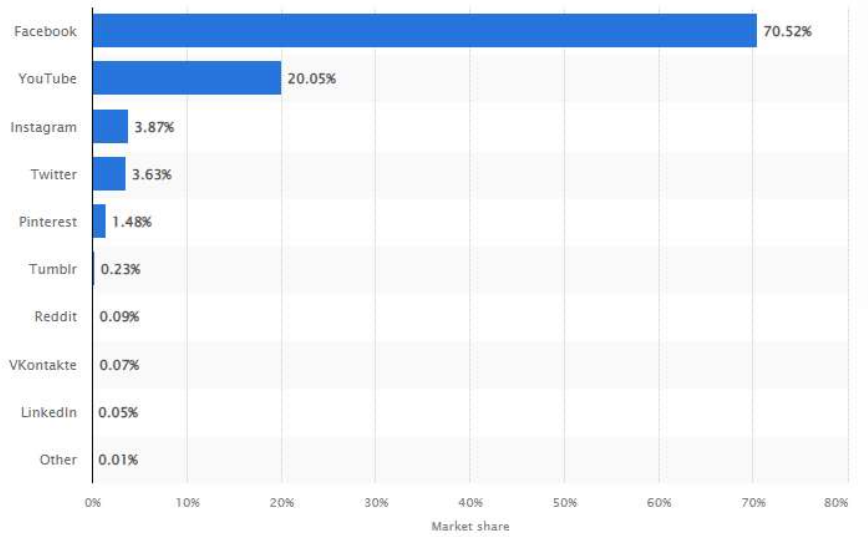


Figure 4 Indonesian Social Media Platform July 2021 chart

Social media is a place for people to communicate, talk, and give opinions about what is going on, especially the Twitter platform, which utilizes some of the world's community as a transparent medium to express opinions about open views (Lubis et al., 2022). Based on the July 2021 Statistics, Indonesian

people's most widely used social media platforms are Facebook, YouTube, Instagram, and Twitter (Hanadian Nurhayati-Wolff, 2021).

The top four most popular social media platforms are Facebook, YouTube, Instagram, and Twitter, as shown in Figure 4. On 24 October 2017, the hashtag #metoo became a trending topic on Twitter which was used to share stories about sexual violence issues and trending topics on Facebook within 24 hours after using the hashtag. Meanwhile, in Indonesia, on 09 June 2021, sexual matters became a trending topic on Twitter, involving a figure board that received a lot of attention and response from the wider Indonesian community. (Lubis et al., 2022)

In previous studies, the SVM Algorithm is the best algorithm from the comparison test results of Naïve Bayes machine learning algorithms, KNN, and SVM with the topic of sentiment analysis in Indonesian (Utami & Masripah, 2021). The SVM algorithm model will be used in this study from the results of the three comparisons of the SVM model, getting the highest accuracy of the other two machine learning algorithms.

There are three frameworks in data mining, namely CRISP-DM, KDD, and SEMMA. Of the three frameworks, CRISP-DM is most suitable for use in various sectors compared to other frameworks, including the IT industry. This research is an IT industry because it performs text mining and sentiment analysis, so CRISP-DM is the right choice and is suitable for this research. CRISP-DM has steps and is well structured (Daderman & Rosander, 2018).

Previous research has conducted comparative experiments of various yahoo word embedding, namely Word2Vec, Glove, and FastText, using two different datasets (20 newsgroups and Routers). From the results of previous studies, the performance of FastText is superior to Glove and Word2Vec because the results of the F1-score obtained by FastText are 0.979 for datasets of 20 newsgroups. In contrast, Routers's dataset gets 0.715 (Nurdin et al., 2020).

Previous studies have done a comparative test of FastText and TFIDF. From the second result, we get the same f1-score, but FastText is superior in terms of time compared to TFIDF, which has a significant difference, namely TFIDF 1.478 seconds while FastText gets 0.0484 seconds (Amalia et al., 2020). From the two previous studies, FastText is superior in word embedding capabilities and fast launch.

Therefore, this study differs from previous research because it has a more specific topic: cases at UMN. Data sources were selected through various social media, namely Twitter, Line Today, Instagram, and Medium. In the flow of this research, FastText is used as word embedding, and the Support Vector Machine or SVM algorithm is used as a machine learning model to predict and analyze sexual sentiment related to UMN.

II. LITERATURE REVIEW

A. Sexual Harassment

Sexual harassment is a sexual act that the victim does not want by an evil person. Harassment is also a form of sexual violence. There are three dimensions of sexuality, namely gender attention, unwanted sexual attention, and sexual coercion. Sexual harassment often occurs in urban areas, campuses, workplaces, offices, or quiet places the victim is a woman. It is not uncommon for men to become victims, but the perpetrators who commit sexual acts are male (Rusyidi et al., 2019).

B. CRISP-DM

CRISP-DM, or Cross-Industry Standard Process for Data Mining, is one of the standardized framework methods for implementing data mining projects. CRISP-DM has six stages or phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment (Schröer et al., 2021).

C. Data Mining

Data Mining is a field of computer science using mathematical algorithms that can collect information on a large scale to find new patterns or information that was not previously known automatically and efficiently and can be used for decision making. Data Mining has been used in various fields such as finance, telecommunications, insurance, and retail sectors, including credit/credit card approval, fraud detection, market segmentation, trend analysis, better marketing, spend analysis, etc. (Bhatia, 2019).

D. Text Mining

Text mining is a particular type of data mining that can mine data from extensive collections of text and extract information from text data to find new and unknown information, such as information in patterns and relationships between texts. Text mining is also used in text classification, text grouping, information extraction, document summarization, and opinion mining or sentiment analysis. Text mining involves natural language processing which can help analyze and process text data, also known as text preprocessing. In addition, text preprocessing can also help with issues related to inconsistent language, inappropriate language use, use of slang, syntax differences, or specific languages. Text mining can help identify information about a sentiment that aims to detect a problem and get a solution (Jo, 2019).

E. Sentiment Analysis

Sentiment analysis identifies and extracts data into information, finds out what they express, and then finds polarity classifications into categories. The categories themselves consist of negative, positive, and neutral. Sentiment analysis can also help understand the sentiment in any situation, such as public sentiment on a condition such as product reviews, financial markets, customer relations, marketing strategies, etc. Sentiment analysis can further analyze text using machine learning and deep learning. With these aspects get more accurate and in-depth. There are three levels of sentiment that can be done. The first are the document, sentence, and aspect levels (Kastrati et al., 2021).

F. Natural Language Processing

Natural language processing is a part of computer science, and artificial intelligence is used to learn or understand a word or language just as humans speak or write a speech. NLP can also study human behavior using language or specific patterns in a data text. Natural language processing is used today because it benefits humans, such as language interpreters, spam detection, location naming, voice settings, chatbots, or bot assistants. Understanding language requires stages because many vocabularies have the same meaning or use concise language or slang. Therefore, natural language processing is carried out. Several stages are carried out during the Natural Language processing process (Lane et al., 2019).

G. Word Embedding

Word Embedding is a word document that can find the meaning of a word with other word relationships by converting a word into a vector of certain words. One of the most effective methods is LSI/LSA. Word Embedding produces a relatively sizeable dimensional vector, and several algorithm methods use word embeddings, such as Word2vec and FastText (Mandelbaum & Shalev, 2016).

H. Fast Text

Fast text is a system that is quite popular among word inserters who can learn word insertion and text classifiers. FastText supports skip-grams with negative sampling and CBOW. FastText is similar to Word2Vec because of the purpose of the vector representation of a word, and FastText is a development of Word2Vec will have its way of achieving the goal. Word2Vec uses words to predict words, whereas FastText has n-gram characters and word insertion obtained by summing n-gram representations (Bojanowski et al., 2017).

I. SVM

The Support Vector Machine or SVM is a machine learning classification and regression algorithm that aims to create boundaries between classes that make it possible to predict the label of more than one vector. The SVM constraint can be called a hyperplane. SVM has kernel functions that can affect the performance of the SVM model for better or worse, such as 'poly,' 'RBF,' 'sigmoid.' (Huang et al., 2018).

J. Confusion Matrix

Confusion Matrix helps provide information and evaluate the results of comparisons made with the original classification. The Confusion Matrix has four values: True Positive, False Positive, False Negative, and True Negative, which are used in a table starting from 2x2 tables to n x n (Diez, 2018).

III. METHODOLOGY

A. Business Understanding

The business understanding stage is the initial stage of CRISP-DM. This stage is to understand and determine the target of the problem in the case of comments related to sexual harassment that occurred at UMN (Karami et al., 2021). Planned targets can create models to predict and analyze the sentiments of sexual harassment comments related to UMN and answer this research's main problem.

B. Data Understanding

Understanding the data will be divided into four sub-processes where
consists of data collection, data labeling, exploratory data analysis, and data cleaning (Schröer et al., 2021).

1. Collecting Data

At the data collection stage, there are two data collection processes where the first data collected is related to sexual regulation in general & not related to UMN. On the other hand, the second Data contains comments about sexual harassment in UMN through social media such as Twitter, Line Today, Medium, and Instagram.

2. Data Labelling

At this stage of the labeling process, two volunteers will be chosen
to help label what sentiment is appropriate with the comments that have been collected. Second, the volunteer is a friend of the author with an A in the language subject Indonesia will help label the sentiment appropriate for each comment from the total available data. There are three sentiment options given to mark existing sentiments three, namely "positive," "neutral," and "negative.".

3. Data Verification

After carrying out the labeling stage, the next step is to choose a linguist to leverage the labels that volunteers have labeled. An expert linguist Indonesian language who will do the leveraging is Niknik Mediyawati, S. Pd., M. Hum, who will bring up and correct the wrong sentiment by giving three highlight colors. Green highlights are marked as positive, red should be harmful, and yellow should be. After that, she will fix the false opinion according to the highlighted color. The linguist will fix the data with a minimum of 100 data, and after that, the label will be fixed manually based on the appropriate sentiment.

4. Exploratory data analysis

The exploratory data analysis stage will perform several visualizations to see the information contained in the data, such as viewing the total data, data size, data statistics, distribution of data distribution, and checking for missing values. To view this information, graphic visualization can be done on certain information so that it is easy to see information, such as using barplot, word cloud, pie chart, etc. (Vadloori & Sanghishetty, 2021).

5. Data Cleaning

After understanding the shape and quality of the data from the results of exploratory data analysis, the next step is to improve the quality of the data by dropping the Data if there are missing values and encoding labels on the dependent variable or predictor, the result of which will be polarity, such as negative sentiment, which will be labeled -1, neutral sentiment will be labeled 0, and positive sentiment will be labeled 1.

C. Data Preparation

Data preparation will be divided into five subprocesses: Preprocessing text, word cloud, word embedding, UMAP, and splitting data (Schröer et al., 2021).

1) Preprocessing Text

At the next stage, where to clear for easy access to the computer. Preprocessing data include Natural Language Processing, from noise deletion, case folding, tokenization, stop word, stemming, and detoxification. The Data that has been saved is then stored in CSV format so that it can be directly used at a later stage or as a checkpoint. There is a condition that if the Data is training data, it will be stored with the name training data and will continue the following process. In contrast, if it is not training data, it will be held under the name UMN sentiment data and go through the procedures contained in the data preparation and go directly to the modeling process.

2) Word Embedding

At the word embedding stage, we will use the FastText method because, in related research, FastText has a pretty good vectorize accuracy and can also vectorize words that have never been encountered before compared to the count vectorizer and TDIDF methods. The research "Comparison of Word Embedding Word2vec, Glove, and FastText in Text Classification" shows that each word embedding algorithm performs similarly and supports the problems and language used. FastText has the advantage of vectorizing a word that does not have a vocab or dictionary, while word2vec and glove have to learn not to vectorize a comment that is not in a dictionary. In addition, the best result from the comparison of the three types of word embedding is FastText (Nurdin et al., 2020). FastText has good results even with Word2Vec in the Indonesian language dataset (Amalia et al., 2020). Therefore, the type of word embedding used in this study is FastText.

3) UMAP

At the UMAP stage, the precise data will be made with the concept of the word bag first, which contains a collection of unique words in an array (Lane et al., 2019). The expression of bags containing these uncommon words will then be vectorized using the Word Embedding technique, such as the word embedding sub-chapter explanation. The vectorized data will have a size of 300x300. Extensive data will be dimensionally reduced using UMAP. The result of the dimension reduction will be n rows and two columns divided into X and Y axes so that the word of bags data can be visualized on the X and Y axes.

4) Slicing Data

Before splitting the data, it will check for missing values from the word embedding process. If the embedding results contain nan or inf, the data will be cleaned by dropping the data. After that, the data will be divided into X and y, where X is independent data and y is predictor data.

5) Splitting Data

The data will be divided into training and testing data at the splitting stage. Based on other studies, it has been proven that using various machine learning algorithms has proven to be the best ratio and is suitable to be used to separate datasets, namely a ratio of 70 to 30 or 70% of training data and 30% of test data (Nguyen et al., 2021). So in this study, the data splitting ratio used is 30% for data testing, while the training data is 70%.

D. Modeling

Machine model training will be carried out at the modeling stage, learning and testing the model results with datasets separated based on the explanation in the literature section (Schröer et al., 2021). The classification technique used is the supervised learning technique because in previous studies comparing the results of the accuracy of supervised learning using linear regression algorithms, decision trees, and SVM to get an average accuracy of 82.33% while unsupervised learning with the K-Means algorithm, a single linkage, and a priori get accurate results with an average of 78%.

After choosing the machine learning technique to be used next, the supervised learning algorithm will be used. In this study, we will use yahoo SVM as a solution to research problems because

based on previous research with the title "Comparison of Classification Algorithms in Sentiment Analysis, Reviews, Online Learning and Distance Education" shows that the results of the average accuracy of SVM get 87.67% better compared to Nave Bayes with an average value of 86.33% using an Indonesian language dataset (Utami & Masripah, 2021). Coupled with other supporting journals that use Indonesian language datasets, it shows that the intermediate results of the SVM research are also better than Naïve Bayes, with a difference of 4.42% (Lutfi et al., 2018).

E. Evaluation
At the evaluation stage, you will see the performance of the prediction results in training datasets and testing datasets using the confusion matrix and visualization metrics and performing calculations from the confusion matrix consisting of precision, recall, and f1-score (Schröer et al., 2021).

F. Deployment
At this stage, the model trained and evaluated will be implemented to predict sentiment data related to sexual harassment at UMN. The results of these predictions will be analyzed using visualization techniques to understand the answers or solutions of this research.

IV. RESULTS AND DISCUSSION

A. Business understanding

Sexual harassment is a sexual act carried out intentionally, and there are indications of coercion against refuse victims (Myrtati D. Artaria, 2012). Based on education level from 2015-2020 in Indonesia, the university ranks first with 14 total cases or 27%. It is followed by Islamic boarding schools, which rank second most significant with ten total cases or 19% (Komisi Nasional Perempuan, 2020).

In addition, Indonesian people widely use social media platforms are Facebook, YouTube, Instagram, and Twitter in July 2021 (Hanadian Nurhayati-Wolff, 2021). However, the #metoo case became a trending topic on Twitter on 24 October 2021. As for Indonesia, on 09 June 2021, it became a trending topic for sexual attention involving a figure on Twitter who got a lot of attention from the Indonesian people (Lubis et al., 2022).

Therefore, this study will use the crisp-dm framework as a research flow and the support vector machine or SVM algorithm as a machine learning model to predict and analyze sexual harassment-related sentiments in UMN through Twitter, Instagram, Line Today, and Medium. The results obtained will be analyzed based on the visualization needed to answer the main problem in this study.

B. Data Understanding

1. Collecting Data
The first data obtained from scraping data on Twitter with the words Indonesian language key relationship is 363 data for the training model, and the second dataset used to conduct sentiment analysis on sexual relations related to UMN is 287 data obtained from various sources such as Twitter, Instagram, Medium, and Line Today.

2. Labelling Data
Figure 5 is the result of labeling sentiment data by two different friends who have sufficient Indonesian language skills because they get an A score.

JOURNAL OF MULTIDISCIPLINARY ISSUES
Journal Website: www.jmis.site
J. Multidisc. Issues 2(2) 1 - 21 (2022)

Figure 5 Labelling Result

3. Labeling Verification Data

Figure 6 is the verification result and the correct sentiment carried out by Niknik Mediyawati, S. Pd. , M. Hum. From the verification results, 91 data are by the sentiment and nine others that are inappropriate, such as line 32, where the labeling is incorrect and justified by giving a red highlight, which means that the sentiment should be negative.

Figure 6 Result of verification labeling

4. Exploratory data analysis

There were 80 who commented positively, 136 who commented neutrally, and 147 who commented negatively.

JOURNAL OF MULTIDISCIPLINARY ISSUES
Journal Website: www.jmis.site
J. Multidisc. Issues 2(2) 1 - 21 (2022)

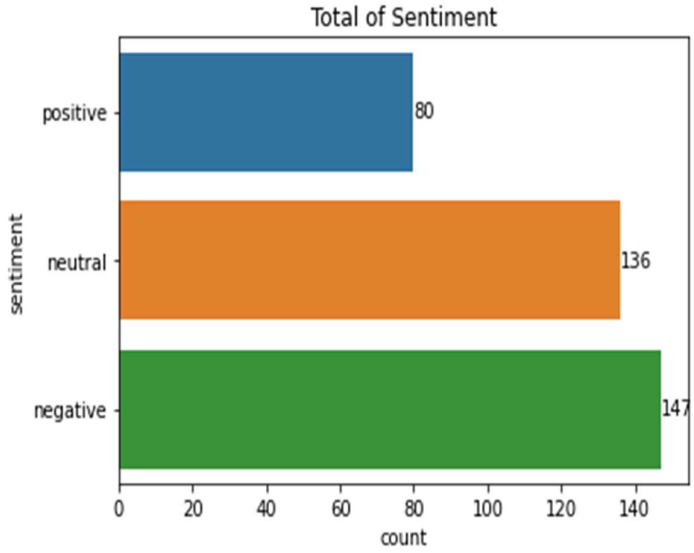


Figure 7 Total sentiment

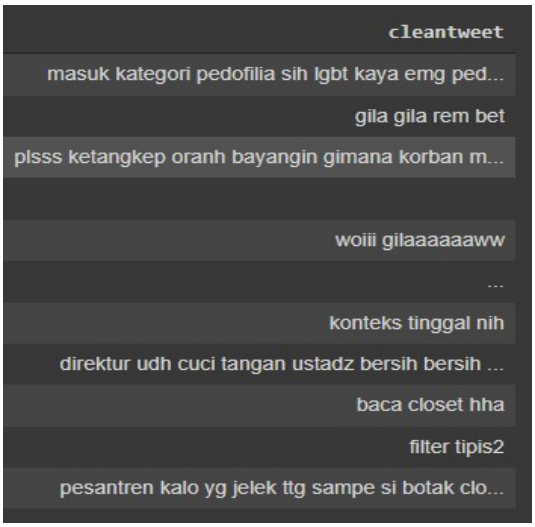
5. Data cleaning
   At this stage, no missing values were found in the data, and the next step was coding the labels. Sentiment label results in the form of text are converted to labeled with numbers. The labeling consists of 3 digits or polarity: label -1 is a negative sentiment label, label 0 is a neutral sentiment label, and label 1 is a positive sentiment label.

tweets	sentiment
ini masuk ke kategori pedofilia sih bukan ke l...	0
Gila gila, serem bet	-1
sss harus ketangkep oranh ini, bayangin gima...	1
https://fin.co.id/read/87804/Viral-Kasus-Dugaa...	1
woiii gilaaaaaaww	-1

Figure 8 Result of label encoding

C. Data Preparation
   1. Preprocessing text
      Several processes flow at the data preprocessing stage, such as missing values, changing target labels, and NLP processes. At this stage, the clean tweet column results from the finished NLP process. This process includes deleting URLs, hashtags, and usernames, changing all text to lowercase text, removing numbers and symbols, tokenizing and eliminating stopwords and stemming, and then doing a detox that combines all previously tokenized words.

JOURNAL OF MULTIDISCIPLINARY ISSUES
Journal Website: www.jmis.site
J. Multidisc. Issues 2(2) 1 - 21 (2022)



cleantweet
masuk kategori pedofilia sih lgbt kaya emg ped...
gila gila rem bet
plsss ketangkep oranh bayangin gimana korban m...
woiii gilaaaaaaww
...
konteks tinggal nih
direktur udh cuci tangan ustadz bersih bersih ...
baca closet hha
filter tipis2
pesantren kalo yg jelek ttg sampe si botak clo...

Figure 9 Result of preprocessing text

2. Word Embedding
The vectorization stage uses the FastText word embedding method to convert each word into a vector that can later be read and studied by machine learning. The result of vectorization is a collection of several vectors combined in 1 array, as shown in Figure 10.

```
Shape:  (363,)
array([array([ 4.94954064e-02, -5.17755048e-03, -2.42018094e-03,  5.01140803e-02,
              -2.52296217e-03, -5.19705079e-02, -6.43790560e-03,  1.39691969e-02,
              -6.08932460e-03, -7.53538981e-02,  1.45116281e-02, -1.49172964e-02,
               1.68084893e-02, -1.24670370e-02,  2.19659302e-02, -2.27778796e-02,
               5.30125154e-03, -8.83121416e-03,  5.40017076e-02,  2.04412313e-03,
              -1.09594045e-02, -4.74794116e-03,  2.13507377e-02,  2.43138988e-03,
               1.23582287e-02,  1.66490823e-02,  1.72245328e-03,  1.28797081e-04,
               2.32517794e-02, -3.75430624e-04,  6.35958016e-02, -2.61271819e-02,
               1.31620392e-02, -4.51412722e-02,  3.22594680e-02,  2.66993511e-02,
               3.08930990e-03, -1.04669863e-02,  1.66079346e-02,  1.82975028e-02,
               2.62035392e-02, -8.24574009e-03, -1.26358727e-02,  9.32180695e-03,
              -3.63446772e-02,  7.70161813e-03, -7.83230830e-03,  2.23946646e-02,
              -2.11532917e-02, -2.84332465e-02, -3.77620161e-02,  2.29456904e-03,
              -5.06859683e-02, -1.52107738e-02, -7.02293823e-04, -6.81480346e-03,
               3.66184022e-03,  1.64660197e-02, -6.35142028e-02, -3.74781080e-02,
              -3.40078734e-02,  8.37469753e-03, -6.35455037e-03,  1.19128264e-02,
```

Figure 10 Result of word embedding

3. UMAP
The UMAP technique can drain the dimensional data into 2 for x and y from 300 columns to visualize data on each word related or similar word. Figure 11 shows the result of the dimension reduction by UMAP broken down into each word with 4598 different words.

JOURNAL OF MULTIDISCIPLINARY ISSUES
Journal Website: www.jmis.site
J. Multidisc. Issues 2(2) 1 - 21 (2022)

	umap1	umap2	text
0	8.291712	0.092955	masuk
1	10.283142	0.236282	kategori
2	9.915542	1.921908	pedofilia
3	6.212286	2.381387	sih
4	8.970923	3.560383	lgbt

Figure 11 UMAP result

For example, Figure 12 shows the visualization of an enlarged image for easy viewing. It can be seen that the collection of words is a word that has negative connotations such as annoyed, annoyed, furious, and restless. Examples of these three words have the same meaning but different words. Therefore, the calculation result of Fast Text vectorization is successful.

Figure 12 Visualization of grouping text

4. Slicing Data
After seeing the vectorization and word relations results, the next step is to check for missing values. It can be seen in Figure 13 that the total available Data is 354 data, of which 9 data do not exist or which are discarded or deleted from the previous data of 363 data.

Missing Values Exist: 9

	0	1	2	3
0	0.049495	-0.005178	-0.002420	0.050114
1	0.062059	-0.002777	-0.005157	0.053669
2	0.031255	-0.033189	0.006001	0.052808
4	0.041624	-0.002576	-0.023213	0.001337
5	0.002524	-0.011105	-0.001025	0.054406
...	...	...	...	...
358	0.011821	-0.006613	0.000044	0.091855
359	0.027229	-0.020452	0.030993	0.069742
360	0.057046	-0.013092	-0.012866	0.012628
361	-0.026860	0.009285	-0.090249	-0.033727
362	0.019945	-0.005096	0.000009	0.074688

354 rows × 300 columns

Figure 13 Result of slicing data and cleaning missing values

5. Splitting Data
The following process is data sharing, where the Data is divided by 30% for test data or as much as 107 data, and the rest is 70% for train data or as much as 247 data with 43 random states.

```
X Train Shape:  (247, 300)
Y Train Shape:  (247, 1)
X Test Shape:  (107, 300)
Y Test Shape:  (107, 1)
```

Figure 14 Result of splitting data

D. Modeling
Figure 15 is the prediction and code from model training using the SVM algorithm or support vector machine using the default parameters.

```
from sklearn.svm import SVC
clf = SVC()
clf = clf.fit(X_train, y_train.values.ravel())
y_pred_train = clf.predict(X_train)
y_pred_train
```

```
array(['-1', '1', '1', '-1', '-1', '1', '-1', '-1', '0', '0', '-1', '0',
       '-1', '1', '-1', '0', '0', '-1', '1', '-1', '1', '-1', '-1', '-1',
       '0', '-1', '-1', '0', '-1', '-1', '-1', '1', '0', '-1', '-1', '1',
       '-1', '-1', '-1', '-1', '0', '0', '-1', '-1', '-1', '0', '-1', '0',
       '0', '1', '-1', '-1', '-1', '1', '-1', '-1', '-1', '0', '0', '1',
       '1', '0', '-1', '-1', '0', '0', '-1', '-1', '-1', '-1', '-1', '0',
       '0', '-1', '0', '0', '-1', '0', '-1', '0', '-1', '-1', '-1', '0',
       '-1', '0', '-1', '0', '0', '1', '0', '-1', '-1', '0', '-1', '-1',
       '-1', '-1', '-1', '1', '0', '-1', '-1', '-1', '0', '0', '0', '-1',
       '1', '-1', '-1', '-1', '0', '-1', '-1', '0', '0', '-1', '-1', '0',
       '1', '1', '1', '0', '0', '1', '0', '1', '0', '0', '0', '0',
       '0', '0', '0', '-1', '0', '0', '0', '0', '-1', '-1', '-1', '-1',
       '0', '-1', '-1', '-1', '0', '0', '-1', '-1', '-1', '-1', '1', '-1',
       '1', '0', '-1', '-1', '0', '1', '1', '-1', '0', '-1', '-1', '-1',
       '1', '-1', '-1', '-1', '-1', '-1', '-1', '-1', '1', '-1', '1', '0',
       '0', '-1', '-1', '0', '-1', '-1', '-1', '-1', '-1', '0', '-1', '1',
       '-1', '-1', '-1', '0', '1', '0', '-1', '1', '0', '-1', '1', '-1',
       '-1', '0', '-1', '0', '-1', '0', '0', '0', '-1', '0', '0', '-1',
       '-1', '-1', '-1', '0', '1', '-1', '0', '0', '1', '-1', '0', '-1',
       '0', '0', '-1', '-1', '-1', '-1', '-1', '0', '0', '0', '-1', '-1',
       '-1', '-1', '0', '-1', '0', '-1', '0'], dtype=object)
```

Figure 15 Result of training model

JOURNAL OF MULTIDISCIPLINARY ISSUES
Journal Website: www.jmis.site
J. Multidisc. Issues 2(2) 1 - 21 (2022)

Figure 16 shows a prediction in the form of a polarity number using the model trained above.

```
y_pred_test = clf.predict(X_test)
y_pred_test
```

```
array(['-1', '0', '1', '-1', '-1', '0', '-1', '0', '0', '0', '-1', '-1',
       '0', '-1', '0', '-1', '0', '-1', '-1', '-1', '-1', '1', '0', '0',
       ' 1', ' 1', '0', '0', '0', '1', '0', '0', ' 1', '0', '0', ' 1',
       '-1', '0', '-1', '0', '-1', '-1', '-1', '-1', '0', '0', '-1', '-1',
       '-1', '-1', '1', '1', '1', '-1', '-1', '-1', '-1', '0', '-1', '-1',
       '0', '0', '-1', '1', '1', '1', '-1', '0', '0', '-1', '1', '-1',
       '0', '-1', '-1', '0', '0', '0', '-1', '-1', '0', '0', '-1', '-1',
       '0', '-1', '-1', '0', '0', '0', '-1', '0', '-1', '-1', '0', '0',
       '0', '0', '0', '-1', '-1', '0', '-1', '0', '-1', '-1', '0'],
      dtype=object)
```

Figure 16 Result of training model

E. Evaluation

At this stage, we will test the performance of the SVM modeling that has been carried out in the previous section. Figure 17 results from the confusion matrix from the training dataset, which got 230 correct predictions from 247 data with 79.35% accuracy.

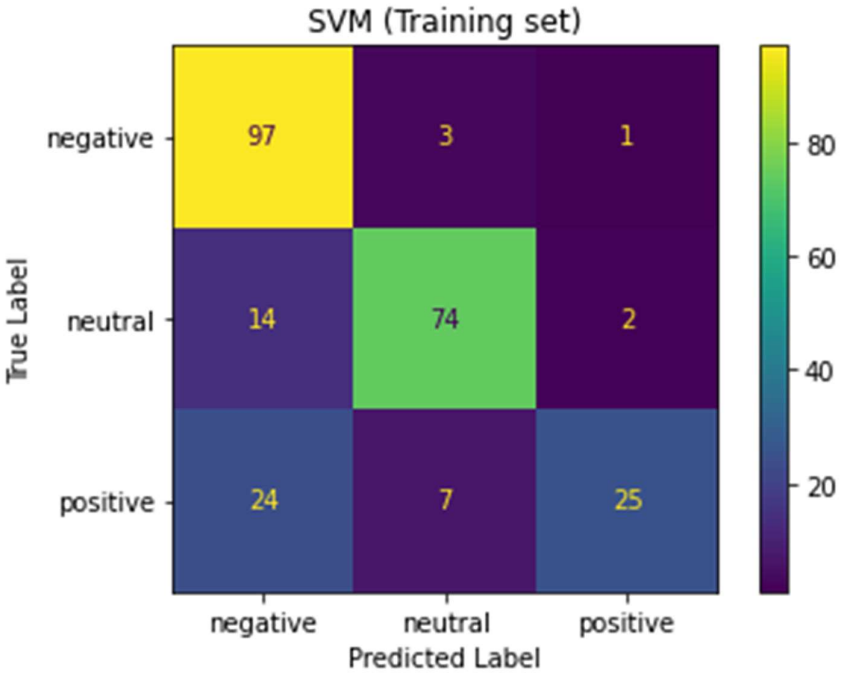


Figure 17 Confusion matrix of training data

Next, evaluate the model that has been tested using a test dataset. Figure 18 is a category confusion matrix of the three sentiments with 67 correct predictions from 107 data with an accuracy of 55.1%.

JOURNAL OF MULTIDISCIPLINARY ISSUES
Journal Website: www.jmis.site
J. Multidisc. Issues 2(2) 1 - 21 (2022)

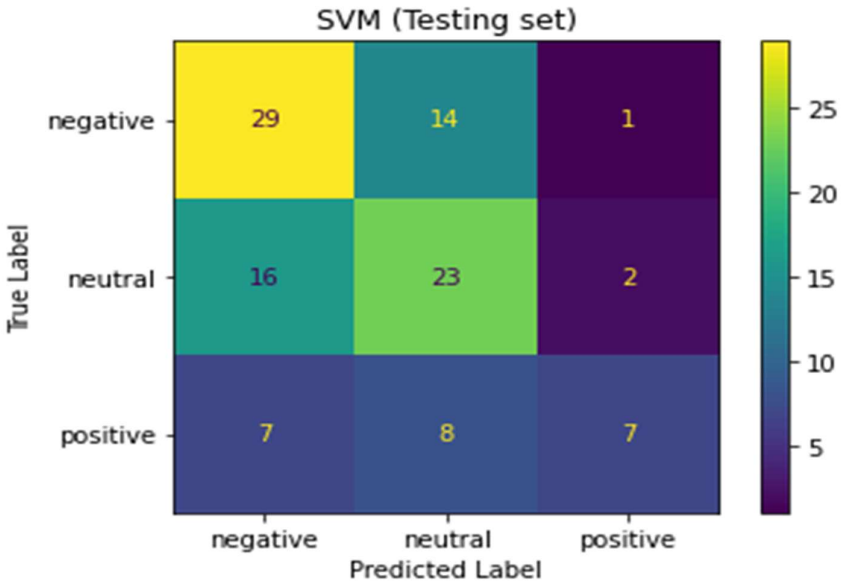


Figure 18 Confusion Matrix of testing data

From the training and testing model, it can be said that it is slightly overfitting because it has a difference of 32.53% from previous research, which got 87.67%. The evaluation results can look from the F1-score of the three labels, which has a relatively sizeable neutral level from the training and testing model results. In contrast, positive sentiment has a lower confidence level than the other two sentiments.

F. Deployment

The deployment stage is the result of the application of the SVM model with a confidence level of 55.14%, which predicts 287 data on sexual harassment comments related to UMN, which can be seen in Figure 19

cleantweet	predict
ckp oke kurang blm tanggap dr otoritas kampus	neutral
huge thanks for speaking up godspeed for your ...	neutral
ur journey will be very long but this is a rea...	neutral
thank you for bringing this topic up to the su...	neutral
duh padal ken banget kuliah jdi takut	negative
...	...
coba tunjukin janji sepakat kampus mahasiswa y...	negative
akun resmi	neutral
bct bgt sih mati aja kali	negative
ngakak min	neutral
lolololololololllll kampus larang larang maha...	negative

Figure 19 Result of prediction by using the SVM model

Based on the predictions of sentiment analysis using Yahoo SVM, which can be seen in Figure 20. Neutral sentiment is predicted to be 54.7%, equivalent to 157 comments from the total data. In comparison, negative sentiment is equal to 36.6% or equivalent to 105 words from the actual data and 8.7% positive sentiment or equivalent to 25 comments which can be seen as follows:

JOURNAL OF MULTIDISCIPLINARY ISSUES
Journal Website: www.jmis.site
J. Multidisc. Issues 2(2) 1 - 21 (2022)

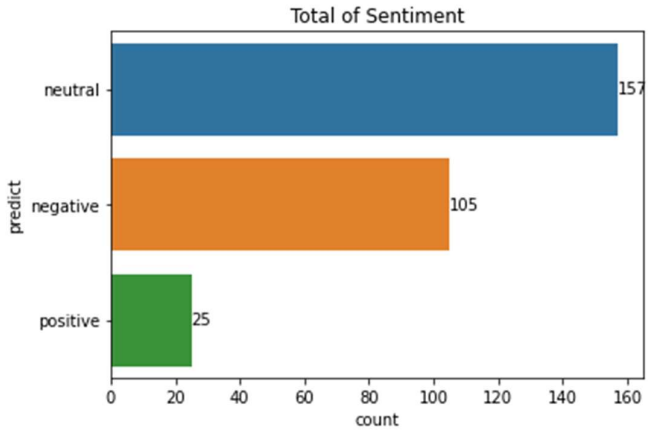


Figure 20 Total sentiment of predicted UMN data

Figure 21 shows that the percentage of data obtained from four sources with the total comments is Twitter at 50.2% and Instagram at 44.6%, while Line Today and Medium are below 4%.

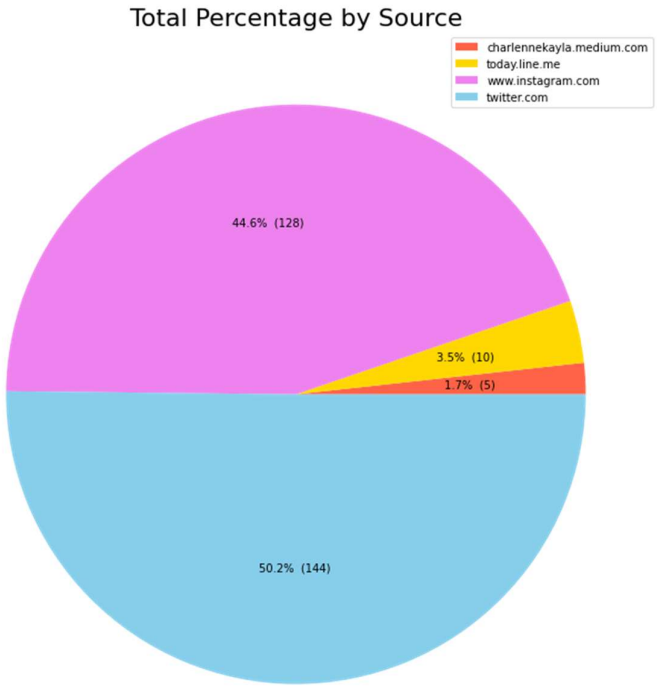


Figure 21 Total percentage based on sources

It can be seen in Figure 22 that Twitter and Instagram are the most prominent because these two sources are the most commonly found comment data. When viewed, the most neutral comments were found from the three sources, namely Medium, Instagram, and Twitter, while the most negative comments were found based on the total frequency of data; LineToday was 70% of the total data and for Twitter only 38.19%, Instagram 32.81 %.

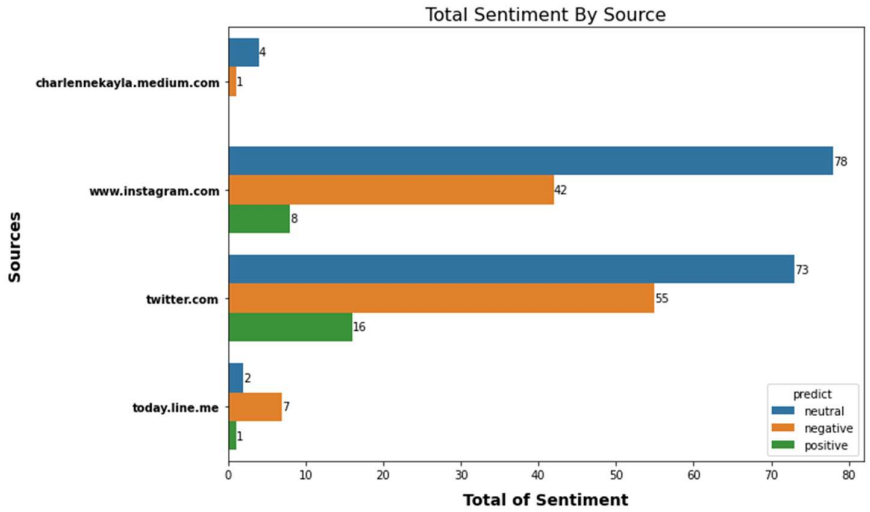


Figure 22 specific total of sentiment based on sources

Figure 23 is a graph of the percentage of comments based on sources and sentiment, which shows that the most negative comments on the Twitter platform are 52.4% or 55 comments, while on Instagram is 40% or 42 comments, but negative sentiment on Line. Seven negative comments out of 10 statements. The most positive sentiment was found on Twitter by 64%, only weighed by 32%, but the Medium platform did not see positive sentimental comments. This is the same as the neutral sentiment of the Instagram platform, which is predicted by most as much as 49.57%, which has a 3.2% difference from Twitter.

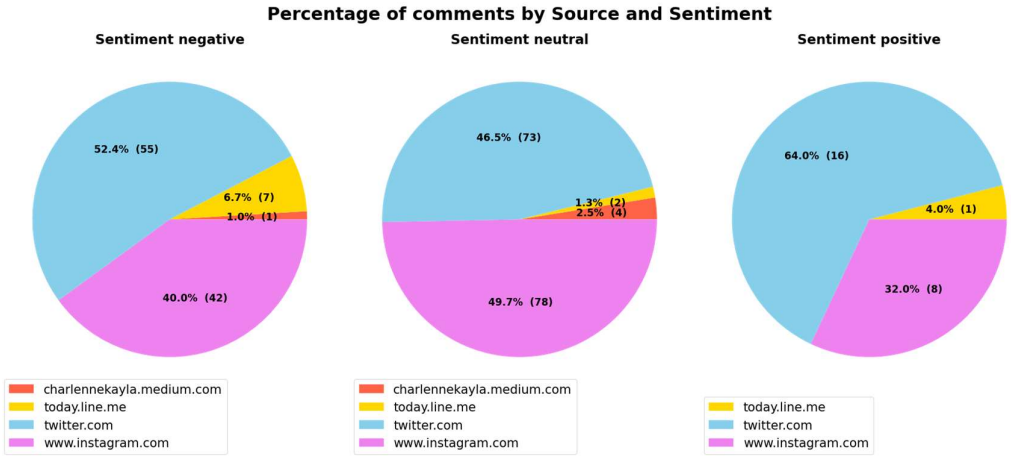


Figure 23 Percentage of Comments by Source and Sentiment Chart

Find words that are often used based on negative sentiments can be seen in Figure 24, showing words that often appear or are used by netizens to comment with the words "kampus," "yg," "leceh," "korban," "kuliah."

JOURNAL OF MULTIDISCIPLINARY ISSUES
Journal Website: www.jmis.site
J. Multidisc. Issues 2(2) 1 - 21 (2022)

Figure 24 Word Cloud of negative sentiment

Figure 25 shows positive sentiments comments; the words that are often used are "kampus", "keren", "respon", "dukung", "berani".

Figure 25 Word Cloud of positive sentiment

Meanwhile, for neutral comments, it can be seen in Figure 26, which shows the word frequently used with a total of 535 words. The word with the highest frequency in neutral comments is "tulis," which is found 16 times in comments with neutral sentiments.

JOURNAL OF MULTIDISCIPLINARY ISSUES
Journal Website: www.jmis.site
J. Multidisc. Issues 2(2) 1 - 21 (2022)

Figure 26 Word Cloud of neutral sentiment

In neutral sentiment comments, some words contain or have positive connotations, such as "terima kasih," "thank," "moga," "good," and "respect," which are found in the top 20 most prominent frequencies. Those words predicted by the machine learning model that should be positive are considered neutral sentiments. Therefore, neutral sentiment will be regarded as a positive and supporting sentiment.

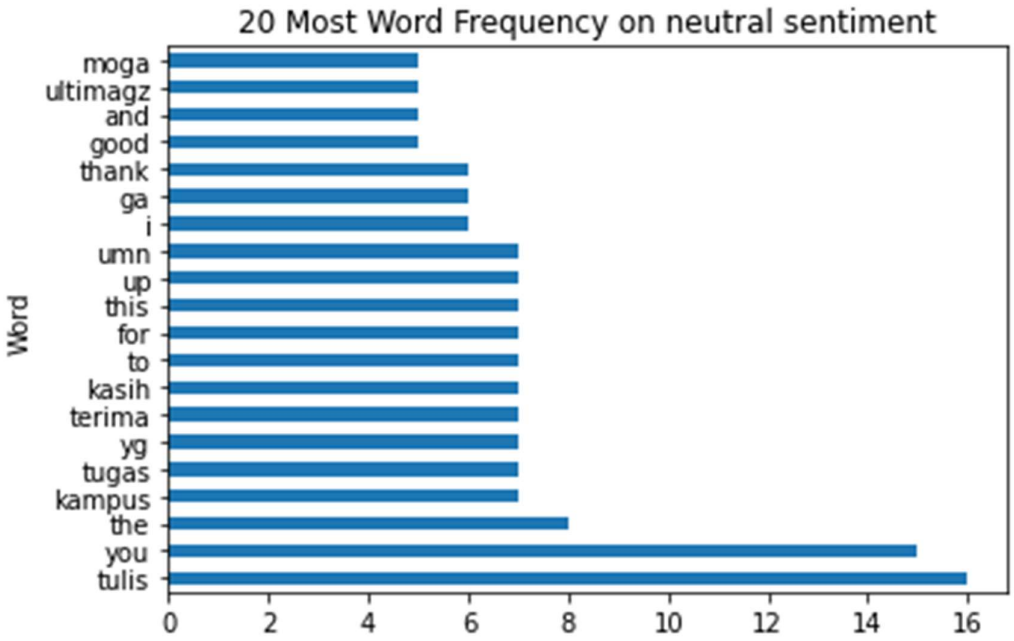


Figure 27 Most words frequency on neutral sentiment

As seen in Figure 27, In neutral sentiment comments, some words contain positive connotations such as "thank you," "thank," "moga," "good," and "respect," which are found in the top 20 most significant frequencies. Those words predicted by the machine learning model that should be

positive are considered neutral sentiments. Therefore, neutral sentiment will be considered positive sentiment as supporting sentiment.

Based on the three pieces of sentiment on word cloud that have described, Survivors can predict the words that will appear in a sexual comment in the UMN campus environment or good things even though some words are inappropriate, such as the word thank you, which should be said to be positive.

V. CONCLUSION

This paper shows how people react to victims of sexual harassment related to the UMN case on social media that use the CRISP-DM framework, FastText, and SVM.

1. Using the SVM algorithm with a high accuracy level of 55.14%, implemented in the sexual harassment dataset related to UMN, found a genuine neutral sentiment of 54.7% or 157 comments, 36.6% or 105 negative sentiments, and 8.7% or 25 positive sentiments. Netizen responses were relatively neutral and supported because comments with neutral sentiments were entirely predictable. Words with positive or supportive connotations were also found in comments with neutral sentiments on cases of sexual harassment around UMN.

2. Based on the SVM algorithm model with an accuracy rate of 55.14%, it was found that the Twitter, Instagram, and Medium platforms get pretty good support when viewed from the frequency of words that appear primarily from neutral and positive sentiments. At the same time, Line Today has many negative sentiment comments from the total data obtained. If one platform is chosen, then Twitter is a platform that gets a good response from Indonesian netizens.

3. After being implemented in cases of sentiment analysis in sexual cases that occurred around UMN with the SVM algorithm, it obtained an accuracy result of 55.14%, while previous studies obtained a relatively high result of 87.67%, which has a significant enough difference of 32,53%. The difference in accuracy results from previous research is on the topic used, namely conducting sentiment analysis on online learning reviews. In contrast, this research is a sentiment analysis on sexual issues related to UMN.

REFERENCES

Amalia, A., Sitompul, O. S., Nababan, E. B., & Mantoro, T. (2020). An Efficient Text Classification Using fastText for Bahasa Indonesia Documents Classification. *2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics, DATABIA 2020 - Proceedings*, 69–75. https://doi.org/10.1109/DATABIA50434.2020.9190447

Bhatia, P. (2019). Data mining and data warehousing : principles and practical techniques. In *Cambridge University Press*. Cambridge Unversity Press.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Daderman, A., & Rosander, S. (2018). Evaluating Frameworks for Implementing Machine Learning in Signal Processing: A Comparative Study of CRISP-DM, SEMMA, and KDD. *Examensarbete Inom Teknik*, 1–36.

Diez, P. (2018). Introduction. In *Smart Wheelchairs and Brain-computer Interfaces: Mobile Assistive Technologies* (Second Edi). Elsevier BV https://doi.org/10.1016/B978-0-12-812892-3.00001-7

Hanadian Nurhayati-Wolff. (2021). *Market share of leading social media platforms in Indonesia as of July 2021*. Statista. https://www.statista.com/statistics/1256213/indonesia-social-media-market-share/

Huang, S., Nianguang, C. A. I., Penzuti Pacheco, P., Narandes, S., Wang, Y., & Wayne, X. U. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics and Proteomics*, *15*(1), 41–42. https://doi.org/10.21873/cgp.20063

Jo, T. (2019). Text Mining: Concepts, Implementation, and Big Data Challenge. In J. Kacprzyk (Ed.), *Springer* (1st ed., Vol. 45). Springer, Cham. https://doi.org/10.1007/978-3-319-91815-0

Karami, A., Spinel, M. Y., White, C. N., Ford, K., & Swan, S. (2021). A systematic literature review of sexual harassment studies with text mining. *Sustainability (Switzerland)*, *13*(12), 1–24. https://doi.org/10.3390/su13126589

Karami, A., Swan, S., & Moraes, M. F. (2020). Space identification of sexual harassment reports with text mining. *Proceedings of the Association for Information Science and Technology*, *57*(1), 1–10. https://doi.org/10.1002/pra2.265

Karami, A., White, C. N., Ford, K., Swan, S., & Yildiz Spinel, M. (2020). Unwanted advances in higher education: Uncovering sexual harassment experiences in academia with text mining. *Information Processing and Management*, *57*(2), 102167. https://doi.org/10.1016/j.ipm.2019.102167

Kastrati, Z., Dalipi, F., Imran, A. S., Nuci, K. P., & Wani, M. A. (2021). Sentiment analysis of students' feedback with NLP and deep learning: A systematic mapping study. *Applied Sciences (Switzerland)*, *11*(9), 3–5. https://doi.org/10.3390/app11093986

Komisi Nasional Perempuan. (2020). Kekerasan Seksual Di Lingkungan Pendidikan. *Https://Komnasperempuan.Go.Id/*, 1–3. https://komnasperempuan.go.id/uploadedFiles/webOld/file/pdf_file/2020/Lembar Fakta KEKERASAN SEKSUAL DI LINGKUNGAN PENDIDIKAN (27 Oktober 2020).pdf

Lane, H., Howard, C., & Hapke, H. M. (2019). *Natural Language Processing in Action(Understanding, analyzing, and generating text with python)*.

Lubis, K., Nisa, I. C., Dalimunthe, P. D., & Perangin-angin, A. B. (2022). Empathy Gap in Social Media Comments for Sexual Harassment Victim. *International Journal : Tradition and Modernity of Humanity*, *2*(1), 26–27. https://talenta.usu.ac.id/tmh

Lutfi, A. A., Permanasari, A. E., & Fauziati, S. (2018). Sentiment Analysis in the Sales Review of Indonesian Marketplace by Utilizing Support Vector Machine. *Journal of Information Systems Engineering and Business Intelligence*, *4*(2), 61–63. https://doi.org/10.20473/jisebi.4.2.169

Mandelbaum, A., & Shalev, A. (2016). *Word Embeddings and Their Use In Sentence Classification Tasks*. http://arxiv.org/abs/1610.08229

Myrtati D. Artaria. (2012). Efek Pelecehan Seksual di Lingkungan Kampus: Studi Preliminer. *BioKultur*, *1*(1), 53.

Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Van Le, H., Tran, V. Q., Prakash, I., & Pham, B. T. (2021). Influence of data splitting on the performance of machine learning models in prediction of shear strength of the soil. *Mathematical Problems in Engineering*, *2021*. https://doi.org/10.1155/2021/4832864

Nurdin, A., Aji, B. A. S., Bustamin, A., & Abidin, Z. (2020). Perbandingan Kinerja Word Embedding Word2Vec , Glove ,. *Jurnal TEKNOKOMPAK*, *14*(2), 77–78.

Rusyidi, B., Bintari, A., & Wibowo, H. (2019). Pengalaman Dan Pengetahuan Tentang Pelecehan Seksual: Studi Awal Di Kalangan Mahasiswa Perguruan Tinggi (Experience and Knowledge on Sexual Harassment: a Preliminary Study Among Indonesian University Students). *Share : Social Work Journal*, *9*(1), 75. https://doi.org/10.24198/share.v9i1.21685

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying the CRISP-DM process model. *Procedia Computer Science*, *181*(2019), 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Utami, L. D., & Masripah, S. (2021). Comparison of Classification Algorithm on Sentiment Analysis of Online Learning Reviews and Distance Education. *Techno Nusa Mandiri: Journal of …*, 106–109. http://ejournal.nusamandiri.ac.id/index.php/techno/article/view/2715%0Ahttp://ejournal.nusamandiri.ac.id/index.php/techno/article/download/2715/880

Vadloori, K. B., & Sanghishetty, S. M. (2021). Exploratory and Sentiment Analysis of Netflix Data. *International Journal of Engineering Research & Technology (IJERT)*, *10*(09), 214–216. https://www.ijert.org/exploratory-and-sentiment-analysis-of-netflix-data

World Health Organization. (2012). Understanding and addressing violence against women. *World Health Organization*, 2–3. https://doi.org/10.1016/B978-0-08-097086-8.35026-7