Journal of Theoretical and Applied Information Technology
30th September 2024. Vol.102. No. 18

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CHI SQUARE FEATURE SELECTION FOR IMPROVING SENTIMENT ANALYSIS OF NEWS DATA PRIVACY TREATS

DEFITROH CHEN SAMI'UN[1], ARIS SUGIHARTO[2], FERRY JIE[3]

[1]Post Graduate School, Master of Information System Diponegoro University Semarang, Indonesia

[2]Departement of Informatics, Diponegoro University, Semarang, Central Java, Indonesia

[3]School of Business and Law, Edith Cowan University, Joondalup WA 6027

E-mail: [1]defitrohsamiun@gmail.com, [2]aris.sugiharto@live.undip.ac.id, [3]F.jie@ecu.edu.au

ID 55335 Submission	Editorial Screening	Conditional Acceptance	Final Revision Acceptance
15-08-24	17-08-2024	12-09-2024	17-09-2024

ABSTRACT

Data security and privacy issues are becoming increasingly pressing in the technology-driven digital era. In 2022, this issue became a major topic in Indonesia and triggered various responses on social media. YouTube, one of the primary platforms, plays a crucial role as a news source. To understand public reactions to this news, sentiment analysis is employed as a research method. The initial stage before conducting sentiment analysis involves data preprocessing, which includes cleaning, case folding, tokenization, slang correction, stemming, and stopword removal. Subsequently, the TF-IDF method is used to assess the significance of words in documents, and Chi-Square feature selection is applied to enhance the performance of the classification model. The main contribution of this study lies in the application of Chi-Square feature selection to improve sentiment analysis accuracy in the context of data privacy threat news. Chi-Square feature selection has proven to be effective in identifying the most relevant features, thereby eliminating irrelevant features and enhancing the accuracy of the classification model. The use of the C5.0 algorithm combined with Chi-Square feature selection achieved the highest accuracy of 87.34%, compared to the 80.14% accuracy achieved without the Chi-Square feature selection method. This research makes a significant contribution by demonstrating that appropriate feature selection methods can substantially improve sentiment analysis model performance, providing a more accurate and effective approach to managing and analyzing sentiment data from social media platforms.

Keywords: *Privacy data, YouTube, Sentiment analysis, Chi-square feature selection, C5.0 algorithm*

1. INTRODUCTION

In today's digital era, characterized by rapid technological advancements, data security and privacy issues have become a major focus due to their significant impact on society. These issues affect the public's understanding of the potential threats arising from insecure data [1]. In 2022, data privacy became a major topic of discussion in Indonesia, sparking various reactions on social media platforms, including YouTube, which is a major news platform. Therefore, understanding how the public responds to news about data privacy is crucial for better management and protection of personal data.

Sentiment analysis plays a key role in understanding public opinions, attitudes, and emotions towards published news [2]. Various classification methods are employed in sentiment analysis to classify sentiments within the data. These methods include Naïve Bayes, Decision Tree, Support Vector Machine (SVM), Random Forest, Artificial Neural Networks (JST), among others.

According to [3], a study was conducted on sentiment analysis on Twitter using the Naïve Bayes algorithm, Decision Tree (C4.5 algorithm), and Random Forest algorithm for LGBT campaigns in Indonesia. The study found that the Naïve Bayes algorithm had the highest accuracy at 83.43%, compared to the Decision Tree and Random Forest algorithms, which had accuracy rates of 82.91%.

The C4.5 algorithm is one of the algorithms frequently used in sentiment analysis. The C5.0 algorithm is an improvement over the C4.5 algorithm [4]. This algorithm has several

advantages, including faster performance, more efficient memory usage, and lower error rates for unseen cases [5].

In sentiment analysis, feature selection methods are employed as preprocessing steps. The aim is to eliminate irrelevant features, reduce computational costs, and enhance the performance of machine learning methods [6]. The Chi-Square method is one of the widely used feature selection methods due to its fast execution time and its ability to improve accuracy by leveraging attributes with the highest significance level [7]. In this study, the Chi-Square method is used to evaluate the level of dependency between features in the dataset and classes [8]. Research by [9], shows that the implementation of the Chi-Square feature selection method in sentiment analysis with the Naïve Bayes algorithm resulted in better accuracy, specifically 93.33%, while sentiment analysis without the feature selection method obtained an accuracy of 73.33%.

This study addresses a notable gap in the field of sentiment analysis by applying the Chi-Square feature selection method to enhance the performance of the C5.0 algorithm, specifically in the context of news related to data privacy threats. Although sentiment analysis techniques have been employed across various domains and platforms, the integration of Chi-Square feature selection with the C5.0 algorithm remains underexplored, particularly concerning discussions of data privacy on social media platforms like YouTube. The research gap this study aims to fill is the insufficient exploration of combining feature selection methods with advanced classification algorithms for sentiment analysis, especially in the evolving context of data privacy concerns. By providing new insights and improving accuracy in sentiment analysis related to data privacy, this study seeks to contribute significantly to both the practical and theoretical understanding of these critical issues.

2. DEFINITION

This research discusses sentiment analysis conducted using the R programming language. The stages of sentiment analysis include data collection of news articles, data preprocessing, data splitting, word weighting, feature selection using the Chi-Square method, classification using the C5.0 algorithm, and evaluation using a confusion matrix [10].

2.1 R Programming Language

R is an open-source statistical programming language that is freely available. This programming language allows for statistical computation and data visualization with enhanced graphics. RStudio is a development environment that facilitates data analysts in importing, exploring, and manipulating data, as well as making statistical calculations and visualizations simpler [11].

2.2 Sentiment Analysis

Sentiment analysis is a method used to identify and classify opinions or comments into two main categories: positive and negative. The primary goal of this process is to understand the emotions or perspectives conveyed in the comments. In practice, this involves using specific techniques to determine whether the comments express positive feelings or dissatisfaction and criticism (negative). In other words, sentiment analysis helps to ascertain whether the content of the comments tends to reflect a positive or negative response about the topic being discussed. [12].

2.3 YouTube

YouTube is a popular video-sharing site on internet that allows people from around the world to watch, upload, and share videos. YouTube also provides a wide range of video types, including news, movies, TV shows, commercials, music, sports events, and more. Currently, YouTube is one of the largest websites in the world, with over 100 million videos watched daily and more than 65,000 new videos uploaded each day [13]. Therefore, in this study, YouTube is used as a source for collecting text-based comment data related to the news "Data Privacy Threats." The comment data was retrieved using a scraping technique with the assistance of the Data Miner extension.

2.4 Pre-Processing

Text preprocessing is the initial step in preparing data before analysis. This stage includes cleaning, case folding, tokenizing, correcting slang words, stemming, and stopword removal [10].

2.5 TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical technique used to evaluate the significance of a word within a collection of documents [1]. Here is the formula for the TF-IDF method as explained in [3].

$$TF - IDF(t,d,D) = TF(t,d) \times IDF(t,D) \quad (1)$$

TF (*t, d*) indicates how often term t appears in document d, while IDF (*t, D*) measures how rare term t is across the entire document collection *D*. Here, *D* refers to the collection of documents, d is a specific document within that collection, and *t* is the term being calculated in the TF-IDF formula.

2.6 Chi Square Feature Selection

Chi-square is a tool to measure the level of statistical relationship between variables. In this study, the chi-square method is used to examine the extent of the relationship between features in the dataset and classes [8]. Here is the chi-square formula according to [14].

$$\chi 2(t,c) = \frac{N(AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)} \quad (2)$$

t : Term
c : Class/Category
N : Total number of training documents
A : Number of documents in category *c* that contain term *t*
B : Number of documents outside category *c* that contain term *t*
C : Number of documents in category *c* that do not contain term *t*
D : Number of documents outside category *c* that do not contain term *t*

In feature selection, it is necessary to calculate the individual Chi-square value for each word against the relevant category, then sort them in descending order. The higher the Chi-square value, the more important the feature is in the classification process [15],

$$X^2\ max(t) = \max_{i=1,\dots,m} \{X^2(t,c_i)\} \quad (3)$$

X^2: statistic indicating the value of X^2 ufor a specific word t in a set of documents
t: Feature being evaluated to determine its significance in distinguishing between classes in text classification
m: Total number of classes in the set of classes used in the analysis
$|c_i|$: Number of specific classes within the set of classes used in text classification analysis
$X^2\ max(t)$: The maximum value$X^2$for a specific word *t* in a set of documents,
$\{X^2(t,c_i)\}$: The X^2 statistic value between the word *t* and a specific class c_i.

2.7 C5.0 Algorithm

The C5.0 algorithm is an advancement of the C4.5 algorithm with more advanced technology to handle inconsistent and incomplete data. C5.0 also introduces boosting methods and reduces the complexity of decision trees to avoid overfitting [4]. C4.5 itself is an enhancement of Iterative Dichotomizer 3 (ID3) and is capable of handling continuous and discrete attributes, data with missing attributes, and considering attributes with different costs [16]. C4.5 also performs pruning on decision trees to remove insignificant branches [5]. The following is the formula for the C5.0 algorithm according to [17] :

$$Entropy\ (S) = \sum_{j=1}^{k} p_j\ log_2(p_j) \quad (4)$$

S : Set of cases
k : Number of classes in the data
j : An index number that indicates every possible value of the attribute of the variable used
|*pj*| : Proportion of *pj* to *S*

After calculating the entropy value of *S*, the next step is to determine the gain value using the following formula:

$$Gain\ (S,A) = Entropy\ (S) - \sum_{i=1}^{m} \frac{|p_j|}{|S|} Entropy(S_i) \quad (5)$$

A : The variable used
m : The number of categories in variable *A*
|*S*| : The number of cases in *S*
|*Si*| : The set of cases in category *i*

Next is the calculation of the gain ratio. The basic formula that can be used for this calculation is as follows:

$$Gain\ Ratio = \frac{Gain\ (S,A)}{\sum_{i=1}^{m} Entropy(S_i)} \quad (6)$$

Gain (S, *A*) : The gain value of a variable
$\sum_{i=1}^{m} Entropy\ (S_i)$: The sum of entropy values within a variable

2.8 K-Fold Cross-Validation

K-fold cross-validation is a method of dividing a sample into *K* groups, where each group is used as validation data (or test data) in turn to evaluate the model's performance. Then, the model's weights are determined by minimizing the sum of squared prediction errors from all groups. This method is easy to use and does not depend on the model's structure, unlike other more complex methods [18].

2.9 Confusion Matrix

A confusion matrix is a tool used to assess the effectiveness of a classification algorithm. Each cell records the number of objects from one actual class that were placed into another class during testing. If objects from a class are placed in the correct class, it indicates that the classification is functioning well. The total number of correctly classified objects is recorded in the main diagonal of the matrix, while non-zero cells outside the main diagonal indicate errors made by the classification algorithm [19].

Here are the key metrics used to assess the performance of a classification model: accuracy, precision, recall, and F1 score [20].

- Accuracy measures the percentage of correct predictions:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (7)$$

- Precision measures the accuracy of positive predictions, which is the ratio of the number of true positive predictions to the total number of positive predictions made.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (8)$$

- Recall or sensitivity assesses how well the model can accurately identify all positive instances.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (9)$$

- F1 Score is the harmonic mean of precision and recall and is useful for handling imbalanced data.

$$F1\ Score = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right) \quad (10)$$

3. RESEARCH METHODOLOGY

Research procedure refers to a series of systematic stages or steps undertaken in the research process. This sequence includes activities from the initial planning phase to the execution and analysis of results. Typically, the research procedure begins with defining the research problem and objectives, followed by data collection through various methods such as surveys, interviews, or experiments. Once the data is gathered, the next step is to analyze it to draw findings and conclusions. This procedure is designed to ensure that the research is conducted in a structured and consistent manner, allowing the researcher to obtain valid and reliable results.

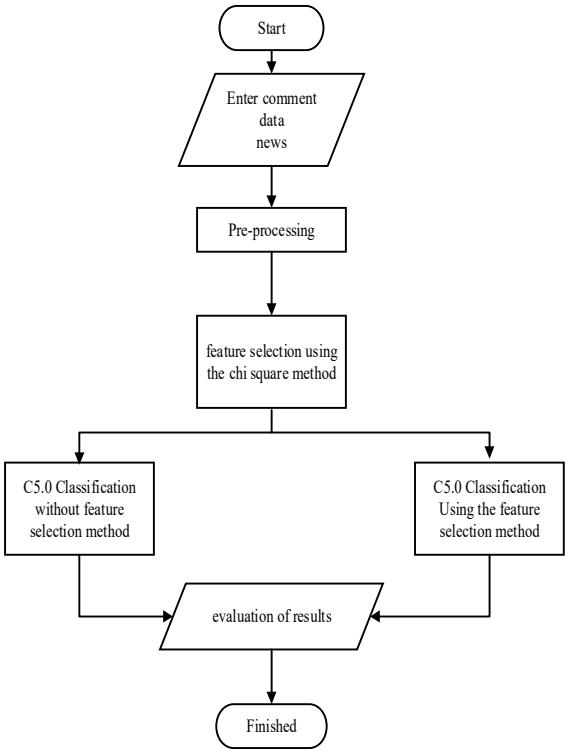

Figure 1: Research Procedure

3.1 Input YouTube Comment Data

Based on Figure 1, the initial step is to input the labeled comment data into the system. The labeling was done manually in Excel by language expert Konstantinus Kapu, S.S, M.Li after extracting the data from YouTube social media. Data extraction of comments from YouTube was conducted using the Data Miner extension. The extracted data consists of comments related to news about data privacy threats.

3.2 Pre-Processing

Next, the normalization process involves several stages, including cleaning, case folding, tokenizing, fixing slang words, stemming, and stopword removal. After the pre-processing stage, calculations will be performed using the Term Frequency-Inverse Document Frequency (TF-IDF) method.

3.3 Feature Selection Using the Chi-Square Method

In this stage, Chi-Square feature selection involves forming a contingency table, calculating the Chi-Square statistic to measure the relationship between features and the target variable, ranking features based on Chi-Square significance, selecting the best features, and testing the model with the selected features. This helps identify features that significantly impact the model's target variable, minimize irrelevant features, and improve model or analysis performance.

In this study, a significance level of $\alpha = 0.05$ is used. Applying $\alpha = 0.05$ in feature selection allows for a more stringent selection of features, which reduces the likelihood of including irrelevant features. This significance level balances the risk of excluding important features with the need for precision, making the process more reliable and effective for subsequent stages of research.

3.4 Classification Using the C5.0 Algorithm

The C5.0 algorithm is a classification algorithm that builds a decision tree model from data divided into training and test sets. The process involves selecting the best features to split the data based on the gain ratio, which measures the increase in information after the split. The resulting decision tree is then pruned to avoid overfitting. The final stage involves validating the model with the test data to evaluate its performance.

3.5 Evaluation of Results

This stage represents the final step or outcome of the sentiment analysis process conducted using the C5.0 algorithm. The analysis involves comparing two different approaches: first, using the Chi-Square feature selection method, and second, without employing any feature selection method. The Chi-Square feature selection method aims to enhance the model's quality by selecting the most relevant and significant features, thereby improving the accuracy of sentiment analysis. In contrast, the approach without feature selection uses all available features without any filtering, which can affect the final results of the analysis. This stage aims to evaluate and compare the effectiveness of both approaches in producing an optimal sentiment analysis model.

4. RESULTS AND DISCUSSION

In this research, comment data sourced from the YouTube social media platform was utilized. The data collection process was facilitated by the Data Miner extension. The data originated from news sources such as CNN, Kompas, CNBC, TvOneNews, and others. The search keywords used were "News Threatening Data Privacy in Indonesia" within the timeframe of 2022-2023. An example of the data used in this study can be seen in Figure 2.

No	Comments	Sentiment
1	Kominfo hrs dihukum krn data nik dan kk sdl	Negatif
2	Maklum . . . | | latar belakang tani koq nanga	Negatif
3	Orang boomer mana tahu teknologi informas	Negatif
4	Gini pak ya, semisal ada pengajuan proyek t	Negatif
5	salah kominfo lah. | | udah tugas nya jaga pri	Negatif
6	Regsitrasi Sim Card ditugaskan Pemerintah l	Negatif
7	Betul setuju perkuat. . . sistem keamananya.	Positif
8	100% salah kominfo lah| | | Tugas hacker en	Negatif
9	SEBENCI apapun kita dengan pemerintah,	Positif
10	Cyber security dah kayak militer sebuah neg	Positif
11	enteng BACOT blg tdk sparah hr ini, , klo tu	Negatif
12	hal ini sangat lucu krn tujuan konminfo minta	Negatif
13	Mau diapain itu data| | Ati ati data buat pemil	Negatif
14	Ini nih yang mesti diperhatikan| | | 1. Masala	Negatif
15	Non aktifkan pejabat di lingkungan kominfo y	Negatif
16	Tau nya cuma blokir situs pekob| | Gk tau yg	Negatif
17	next, Bocor Data My Pertamina| , nanti sal	Negatif
18	Itulah makanya pak bos, layanan nyimpen d	Negatif

Figure 2: Sample Data Used

4.1 Pre-Processing

Text preprocessing is the initial step in preparing data before analysis. This stage includes cleaning, case folding, tokenizing, fixing slang words, stemming, and stopword removal.

- **Cleaning**
  In this stage, comment data containing numbers, URLs, emojis, and characters other than letters will be removed [2].
- **Case folding**
  All letters in the text are converted to lowercase [21].
- **Tokenizing**
  This stage involves breaking down long texts into words or phrases. The goal is to create a

list of words and count how many times those words appear in the text [2].

- **Fixing slang words**
  This stage involves the process of converting non-standard words into standard words or converting abbreviations or informal words into their actual meanings [22].
- **Stemming**
  Stemming is a technique used to convert words in a text to their base or root form [2]. This research utilizes the Nazief and Adriani Stemming method.

  Nazief and Adriani Stemming is a library used to convert words in text to their base form, using morphological rules, affixes, and dictionaries of base words, to make the text easier to understand and process [23].
- **Stopword removal**
  This stage is the final preprocessing step before proceeding to the next analysis stage. In this stage, unimportant words that only serve as connectors in sentences without affecting sentiment will be removed [2]. The results of the data preprocessing can be seen in Figure 3.

No	Comments	Sentiment
1	kominfo hrs hukum data nik kakak bocor buanyak telpon w...	Negatif
2	maklum latar belakang tani koq nanganin kominfo bocor la...	Negatif
3	orang boomer tahu teknologi informasi hilang kominfo mas...	Negatif
4	gin pak misal aju proyek aman siber tuji aja pak rakyat bapa...	Negatif
5	kominfo lah udah tugas nya jaga privasi data orang indones...	Negatif
6	regsitrasi sim card tugas pemerintah kominfo daftar sim car...	Negatif
7	betul tuju kuat sistem keamananya rekrut orang ahli baik m...	Positif
8	kominfo lah tugas hacker emang gitu misal ngehack shopi h...	Negatif
9	benci apa pemerintah bjorka musuh ungkap buruk pemerin...	Positif
10	cyber security dah kayak militer negara kasus ajar harga ind...	Positif
11	enteng bacot bilang tidak sparah hr tu daata dmanfaatkn ut...	Negatif
12	lucu tuju konminfo minta data utk lindung ngtif penyalahgu...	Negatif
13	diapain data ati ati data milu	Negatif
14	nih mesti hati masalah bocor data biasa tidak ajar bocor dat...	Negatif
15	non aktif jabat lingkung kominfo kait aman database guna l...	Negatif
16	tau nya blokir situs pekob tidak tau gin	Negatif
17	next bocor data my pertamina an tuh	Negatif
18	pak bos layan nyimpen data spt google meta kemarin tidak ...	Negatif

Figure 3: Data Pre-Processing Results

4.2 Splitting Data

In this research, the data comprising 404 samples will be divided using the 10-fold cross-validation technique. The data is randomly split into 10 subsets (folds). In each iteration, one-fold is used as the test set, while the remaining nine folds are used for training. This process is repeated 10 times, so each fold serves as the test set once and as part of the training data nine times. The evaluation results from each iteration are averaged to provide a more accurate and representative estimate of the model's performance. Although using $k = 10$ increases computational complexity and the time required, this method provides a more comprehensive and reliable evaluation of the model.

4.3 Application of the TF-IDF Method

After the data has undergone preprocessing, the next step is to process it using the TF-IDF (Term Frequency-Inverse Document Frequency) method. This method aims to assess the importance of words within a document relative to the entire collection of documents. The process begins by calculating the term frequency (TF), which measures how often a word appears in a specific document. Next, the inverse document frequency (IDF) is evaluated, which indicates how rare or common the word is across all documents in the collection. Finally, the TF and IDF values are combined to determine the final score of each word, reflecting its significance to the document. The results of applying the TF-IDF method can be seen in Figure 4.

	bocor	buanyak	data	gilir	hrs	hukum	jelas	kakak	kominfo	laku	nik	nomor	salah	telpon	tidak	whatsapp	blokir
1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1
4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 4: Results Of Applying The TF-IDF Method

4.4 Application of the Chi-Square Feature Selection Method

The process of applying the chi-square feature selection method starts with creating a contingency table that displays the values of words in relation to different categories. This table helps in determining the frequency of each word occurring within each category. Following this, the chi-square value is computed for each word by comparing its observed frequency to the expected frequency in the given category. These chi-square values are then sorted in descending order to identify the most significant features. The results, including the p-

values derived from the chi-square calculations, are used to assess the relevance of each word in the context of the categories analyzed. This procedure helps in selecting the most influential features for improving the performance of the classification model, as illustrated in Figure 5.

word	p_value
anak	0.0339830085
atas	0.1309345327
bank	0.5222388806
bri	1.0000000000
buka	1.0000000000
ceramahin	1.0000000000
emak	1.0000000000
hacker	0.5632183908
idiot	1.0000000000
katain	1.0000000000
kaya	0.6011994003
kuat	0.4402798601
marahin	1.0000000000
nama	0.6206896552
nya	0.6581709145
pantes	1.0000000000
pas	1.0000000000
sih	0.5827086457
sistem	0.0204897551
tarif	1.0000000000
tdi	1.0000000000

Figure 5: P-value results from the Chi square method

4.5 Evaluation of Results

Testing is carried out using the k-fold cross-validation method by dividing the data into 10 parts. Out of these 10 parts, 1 part is used as the test data, while the remaining 9 parts are used for training. This process is repeated 10 times to ensure that each part of the data serves as the test data once. K-fold cross-validation is a model evaluation technique where data is divided into K groups, and each group is used alternately as a validation sample to assess the model's performance [5]. Meanwhile, the confusion matrix is a table used to evaluate the performance of classification algorithms [24]. In accuracy calculations, a 2 x 2 confusion matrix is used to obtain accuracy, precision, and recall values for each sentiment class.

The implementation of the C5.0 algorithm involves data preprocessing, data splitting, feature selection using the Chi-square method, and converting text into a bag-of-words representation. The C5.0 model is then trained and evaluated, with the average metrics from cross-validation providing a comprehensive assessment of the model's performance in sentiment classification. The results of applying the C5.0 algorithm with Chi-square feature selection and without Chi-square feature selection are shown in Table 1.

Table 1: Results

Results	C5.0	C5.0 + Chi-Square Feature Selection
Accuracy	80.14%	87.34%
Precision	50%	78%
Recall	40%	82%

Based on Table 1, the highest accuracy is achieved with the C5.0 algorithm using the Chi-square feature selection method, which is 87.34%. This indicates that integrating Chi-Square feature selection with the C5.0 algorithm significantly improves performance. The testing reveals that Chi-Square feature selection enhances the model's ability to classify sentiment accurately, addressing a major research gap by optimizing advanced classification algorithms for sentiment analysis. The increased accuracy, precision, and recall demonstrate the effectiveness of this approach, consistent with the evaluation criteria used. In comparison, other sentiment analysis methods show varying performance, with C5.0 + Chi-Square outperforming in accuracy.

Table 2: Several Studies Related to Sentiment Analysis

No	research by	Method	Dataset	Accuracy	F1-score
1	This research	C5.0 + Chi Square	YouTube comment data	87.34%	79%
2	[25]	SVM	Qualitative feedback data from students of the University of Education, Winneba	63.79%	63%
3	[26]	KNN	Tweet data about Moderna vaccine	80%	88%
4	[27]	Naïve Bayes	Social media tweet data	85%	77%
5	[28]	Decision Tree	User review data on the PeduliLindungi application	86%	85%
6	[29]	Random Forest	PIMA Indians dataset	75%	80%

The results in Table 2 illustrate the varying performance of different sentiment analysis and classification methods across different datasets. The C5.0 algorithm, when combined with the Chi-Square feature selection method, achieved the highest accuracy of 87.34%, indicating its effectiveness in classifying YouTube comments. Conversely, K-Nearest Neighbor (KNN) excels in F1-score with a value of 88%, reflecting an excellent balance between precision and recall on the Moderna vaccine tweet data. The Decision Tree also demonstrated solid performance with an accuracy of 86% and an F1-score of 85%, highlighting its capability in analyzing PeduliLindungi app reviews. On the other hand, Support Vector Machine (SVM) showed lower results, with an accuracy of 63.79% and an F1-score of 63%, indicating limitations in data classification. Random Forest and Naïve Bayes, while providing solid F1-scores, exhibited lower accuracy compared to some other methods, suggesting that each method has its strengths and weaknesses depending on the type of data analyzed.

4. CONCLUSION

This study specifically focuses on improving sentiment analysis performance related to news on data privacy threats through the application of the Chi-Square feature selection method. In this context, the C5.0 algorithm was used to classify positive and negative sentiments from 404 comment samples collected from the YouTube platform. Significant features were selected based on a p-value of less than 0.05 in the Chi-Square selection process, ensuring that only the most relevant features were used in the analysis.

The study's results indicate that the use of the Chi-Square feature selection method significantly improved the accuracy of the C5.0 model. The C5.0 algorithm with Chi-Square feature selection achieved an accuracy of 87.34%, higher than the 80.14% accuracy obtained from the algorithm without feature selection. This increase in accuracy suggests that the Chi-Square feature selection method is effective in enhancing sentiment analysis performance by selecting the most influential features in determining sentiment related to data privacy threats.

RECOMMENDATION

For future research, it is recommended that researchers expand the scope of analysis by collecting data from various social media platforms, not just YouTube, to enhance the model's generalizability. Additionally, employing other classification algorithms such as Random Forest and Support Vector Machine could provide further insights into the effectiveness of different methods. The research should also consider addressing data distribution imbalances by using smoothing methods.

REFERENCES:

[1] M. Das, S. K., and P. J. A. Alphonse, "A Comparative Study on TF-IDF feature Weighting Method and its Analysis using Unstructured Dataset," Aug. 08, 2023.

[2] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014.

[3] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Sentiment Analysis of Social Media Twitter with Case of Anti-LGBT Campaign in Indonesia using Naïve Bayes, Decision Tree, and Random Forest Algorithm," *Procedia Computer Science*, vol. 161, pp. 765–772, Jan. 2019.

[4] S. Pang and J. Gong, "C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks," *Systems Engineering - Theory & Practice*, vol. 29, no. 12, pp. 94–104, Dec. 2009.

[5] A. S. Galathiya, A. Ganatra, and C. Bhensdadia, "Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning," 2012. Accessed: Oct. 29, 2023.

[6] J. Hu, K. Pan, Y. Song, G. Wei, and C. Shen, "An improved feature selection method for classification on incomplete data: Non-negative latent factor-incorporated duplicate MIC," *Expert Systems with Applications*, vol. 212, p. 118654, Feb. 2023.

[7] L. A. C. Ahakonye, C. I. Nwakanma, J.-M. Lee, and D.-S. Kim, "SCADA intrusion detection scheme exploiting the fusion of modified decision tree and Chi-square feature selection," *Internet of Things*, vol. 21, p. 100676, Apr. 2023.

[8] S. A. Ali *et al.*, "An Optimally Configured and Improved Deep Belief Network (OCI-DBN) Approach for Heart Disease Prediction Based on Ruzzo–Tompa and Stacked Genetic

Algorithm," *IEEE Access*, vol. 8, pp. 65947–65958, 2020.

[9] Nurhayati, A. E. Putra, L. K. Wardhani, and Busman, "Chi-Square Feature Selection Effect on Naive Bayes Classifier Algorithm Performance for Sentiment Analysis Document," in *2019 7th International Conference on Cyber and IT Service Management (CITSM)*, Nov. 2019, pp. 1–7.

[10] E. B. Santoso, Y. H. Chrisnanto, and G. Abdillah, "Identification of Hoax News in the Using Community TF-RF and C5. 0 Tree Decision Algorithm," *Enrichment: Journal of Multidisciplinary Research and Development*, vol. 1, no. 6, pp. 336–351, 2023.

[11] S. C. Statistics, "The Quantitative Analysis of Natural Populations: Some Common Statistics and What They Mean," 2018, Accessed: Nov. 01, 2023.

[12] S. M. Learning, "Hybrid model for twitter data sentiment analysis based on ensemble of dictionarybased classifier and stacked machine learning classifiers-svm, knn and c5. 0," *J. Theor. Appl. Inf. Technol.*, vol. 98, no. 04, 2020, Accessed: Oct. 30, 2023.

[13] D. O'Brien and B. Fitzgerald, "Digital copyright law in a YouTube world," *Internet Law Bulletin*, vol. 9, no. 6 & 7, pp. 71–74, 2006.

[14] F. Thabtah, M. A. H. Eljinini, M. Zamzeer, and W. M. Hadi, "Naïve Bayesian Based on Chi Square to Categorize Arabic Data," *Communications of the IBIMA*, vol. 10, 2009.

[15] P. R. B. Putra, I. Indriati, and R. S. Perdana, "Klasifikasi Judul Berita Online menggunakan Metode Support Vector Machine (SVM) dengan Seleksi Fitur Chi-square," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 5, Art. no. 5, Aug. 2023.

[16] A. D. Mankar and S. D. Bhoite, "A Comparative Study of Recursive Partitioning Algorithms (ID3, CART, C5. 0) for Classification," Feb. 2023, doi: 032021-11278686.

[17] R. Pratiwi, M. N. Hayati, And S. Prangga, "Perbandingan Klasifikasi Algoritma C5.0 Dengan Classification and Regression Tree (Studi Kasus: Data Sosial Kepala Keluarga Masyarakat Desa Teluk Baru Kecamatan Muara Ancalong Tahun 2019)," *Barekeng*, Vol. 14, No. 2, Pp. 273–284, Sep. 2020.

[18] X. Zhang and C.-A. Liu, "Model averaging prediction by K-fold cross-validation," *Journal of Econometrics*, vol. 235, no. 1, pp. 280–301, Jul. 2023.

[19] R. Susmaga, "Confusion Matrix Visualization," in *Intelligent Information Processing and Web Mining*, M. A. Kłopotek, S. T. Wierzchoń, and K. Trojanowski, Eds., in Advances in Soft Computing. Berlin, Heidelberg: Springer, 2004, pp. 107–116.

[20] Azwarni And N. Shah, "Evaluating Textblob, Lexicon, Support Vector Machine, Naive Bayes, And Chatgpt Approaches for Sentiment Analysis of Nasdaq Listed Companies," *Vol.*, Vol. 102, No. 13, Jul. 2024.

[21] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving text preprocessing for student complaint document classification using sastrawi," in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, 2020, p. 012017.

[22] U. Naseem, I. Razzak, and P. W. Eklund, "A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter," *Multimed Tools Appl*, vol. 80, no. 28, pp. 35239–35266, Nov. 2021.

[23] J. Asian, H. E. Williams, and S. M. Tahaghoghi, "Stemming indonesian," in *Proceedings of the Twenty-eighth Australasian conference on Computer Science-Volume 38*, Citeseer, 2005, pp. 307–314. Accessed: Nov. 01, 2023.

[24] O. Alsemaree, A. S. Alam, S. Gill, and S. Uhlig, "Sentiment analysis of Arabic social media texts: A machine learning approach to deciphering customer perceptions," *Heliyon*, p. e27863, Mar. 2024.

[25] D. K. Dake and E. Gyimah, "Using sentiment analysis to evaluate qualitative students' responses," *Educ Inf Technol*, vol. 28, no. 4, pp. 4629–4647, Apr. 2023.

[26] M. I. Hutapea and A. P. Silalahi, "Moderna's Vaccine Using the K-Nearest Neighbor (KNN) Method: An Analysis of Community Sentiment on Twitter," *Jurnal Penelitian Pendidikan IPA*, vol. 9, no. 5, pp. 3808–3814, 2023.

[27] F. Y. Dharta, A. J. Mahardhani, S. R. Yahya, A. Dirsa, and E. M. Usulu, "Application of Naive Bayes Classifier Method to Analyze Social Media User Sentiment Towards the Presidential Election Phase," *Jurnal Informasi dan Teknologi*, pp. 176–181, 2024.

[28] C. M. S. Ramdani, A. N. Rachman, and R. Setiawan, "Comparison of the Multinomial

Naive Bayes Algorithm and Decision Tree with the Application of AdaBoost in Sentiment Analysis Reviews PeduliLindungi Application," *IJISTECH (International Journal of Information System and Technology)*, vol. 6, no. 4, Art. no. 4, Dec. 2022.

[29] F. Mustofa, A. N. Safriandono, A. R. Muslikh, and D. R. I. M. Setiadi, "Dataset and feature analysis for Diabetes Mellitus classification using random forest," *Journal of Computing Theories and Applications (JCTA)*, vol. 1, no. 1, pp. 41–49, 2023.