INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Development of Extraction Features for Detecting Adolescent Personality with Machine Learning Algorithms

Irzal Arief Wisky [a,*], Sarjon Defit [b], Gunadi Widi Nurcahyo [b]

[a] *Department of Information System, University of Putra Indonesia YPTK, Lubuk Begalung, Padang, Indonesia*
[b] *Department of Information Technology, University of Putra Indonesia YPTK, Lubuk Begalung, Padang, Indonesia*
Corresponding author: [*]irzal.arief12@gmail.com

***Abstract*—This study aims to develop a Natural Language Processing (NLP)-based feature extraction algorithm optimized for personality type classification in adolescents. The algorithm used is TF-IDF + N-Gram Z, which combines Term Frequency-Inverse Document Frequency (TF-IDF) with the N-Gram Z technique to improve the feature representation of the analyzed text. TF-IDF functions to measure the importance of words in a document, while N-Gram Z enriches the context by considering the order of words that appear sequentially. The dataset in this study consists of 3,200 sentences generated by adolescent respondents through a survey designed to explore aspects of their personality. After the feature extraction process is complete, three variants of the Naïve Bayes method are applied for classification, namely Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Complement Naïve Bayes. Each variant has distinctive characteristics in handling certain data types, such as binomial and multinomial data. The results of the study show that the combined TF-IDF + N-Gram Z algorithm can produce highly representative features, as evidenced by high classification performance. The Multinomial Naïve Bayes and Complement Naïve Bayes variants each achieved 98% accuracy. These findings provide significant contributions to the development of NLP-based personality classification methods for Detecting Adolescent Personality. The combination of the TF-IDF + N-Gram Z algorithm with various Naïve Bayes variants produces an exceedingly high level of accuracy and can be applied in practice in the fields of psychology and adolescent education.**

***Keywords*— Natural language processing; TF-IDF+N-Gram Z; detecting adolescent personality; naïve bayes.**

Manuscript received 4 Jun. 2024; revised 20 Aug. 2024; accepted 21 Oct. 2024. Date of publication 30 Nov. 2024.

I. Introduction

Feature extraction is the process of identifying feature values in documents for text mining [1]. This step is essential in document processing for search engines, as it significantly influences the success of the text mining process [2]. Feature extraction involves the operation of word weighting [3], which aims to assign values or weights to terms present in a document [4]. Machine learning employs various methods for word weighting, including TF-IDF, TR-RF, WIDF, and Bag of Words (BoW), among others [5], [6]. TF-IDF is particularly favored due to its notable impact on improving accuracy [7], [8], [9], [10], and [11]. It is extensively utilized in text analysis tasks such as sentiment analysis and spam detection [12], [13], and [14].

Several studies utilizing TF-IDF for text analysis include research using the LinearSVC algorithm, which achieved an accuracy of 89% [15]. Another study applied TF-IDF for sentiment analysis using the pseudo-nearest neighbor (PNN) algorithm, achieving a maximum accuracy of 92% [16]. Additionally, researchers often combine TF-IDF with N-Gram techniques, which represent sequences of n units, typically single characters or strings separated by spaces. N-Grams are segments of n-characters extracted from a string, with blanks added at the start and end to delineate string boundaries. The advantage of N-Grams in string matching lies in their ability to mitigate errors, as a mistake in part of the string only affects a portion of the N-Grams.

Researchers have conducted research combining TF-IDF and N-Gram to detect fake news on Twitter using the Support Vector Machine (SVM) algorithm, achieving 90% accuracy [17]. This research processed TF-IDF first, followed by N-Gram. In contrast, the present study adopts a slightly different approach by processing TF-IDF and N-Gram simultaneously, which enhances the accuracy of the machine learning model employed.

For testing the combination of TF-IDF and N-Gram, this study utilizes data from questionnaires to determine

personality types. An individual's personality type significantly influences their character, attitude, and behavior. Swiss psychologist Carl Gustav Jung classified human personalities into three types: introvert, extrovert, and ambivert [18]. Primarily focused on their inner thoughts, introverts tend to be quieter or more reflective [19]. Extroverts derive satisfaction from external sources, enjoy human interaction, and are usually enthusiastic, talkative, assertive, and sociable [20]. Ambiverts fall between these two extremes, comfortable with social interactions but also enjoying solitary time [21].

Several researchers have detected personality using machine learning, employing datasets comprising images or text. Hernandez et al. (2022) classified social media text to determine personality types such as dominance, influence, steadiness, or compliance using various machine learning algorithms like Naïve Bayes, SMO, KNN, AdaBoost, J48, and Random Forest, achieving the highest accuracy of 78% with AdaBoost [22]. Another study detected personality on Twitter, classifying it into labels such as openness, agreeableness, neuroticism, conscientiousness, and extroversion, using XGBoost and achieving 75% accuracy [23]. Additionally, another study used signatures to determine personality, identifying signature types such as curved backward, sharply curved, increasing, and decreasing. The SVM+Feature Extraction PCA algorithm yielded an accuracy of 69.44% [24].

Previous research suggests that we can view personality from various perspectives. Following this, preprocessing was performed to ensure that the data was free from noise, which can lead to low accuracy [25]. To enhance the accuracy of detection, this study employed RAKE during the preprocessing stage. RAKE (Rapid Automatic Keyword Extraction) is an algorithm designed to automatically extract keywords or important phrases from a text [26]. This algorithm identifies keywords based on linguistic characteristics such as word frequency, the presence of words in relevant contexts, and specific word placement patterns [27].

Word weighting in this study was performed using a combination of TF-IDF and N-Gram. The N-Gram features utilized include Unigram, Bigram, and Trigram. The results of the word weighting were further processed using a family of Bayesian algorithms. Three Bayesian algorithms were used: Multinomial Naïve Bayes (MNB), Bernoulli Naïve Bayes (BNB), and Complement Naïve Bayes (CNB). Additionally, this study conducted tests using the previous TF-IDF combination to compare the accuracy results from previous research with the advancements made in this study.

II. Material and Method

A. Literature Review

This research uses several experiments that will be carried out using multinomial naïve bayes (MNB), Bernoulli Naïve Bayes (BNB), and Complement Naïv Bayes algorithms. (CNB). Table I is a model that has been done by previous research and development to be done.

TABLE I
PERSONALITY TYPE CLASSIFICATION QUESTIONNAIRE

No	Preprocessing	Feature Extraction	Algorithm	Ref.
1	Tokenization, stop word removal, removal of punctuations etc.	TF-IDF + N-Gram	Multinomial Bayes	[28]
2	Data Cleaning, Case Folding, Tokenizing, Normalization, stop word removal, and stemming		Bernoulli Bayes, Complement Bayes	[29]
3	Tokenizing & Punctuation, stop word Removal, and Stemming	BoW + N-Gram	Multinomial Bayes	[30]
4	Removing stop words, stemming, lemmatizing, lowercasing, and tokenize		Bernoulli Bayes, Complement Bayes	[31]
5	Data Cleaning, Case Folding, RAKE, Stemming	TF-IDF + N-Gram	Multinomial Bayes, Bernoulli Bayes, Complement Bayes	This Research

From several previous studies, this research has differences in the preprocessing used, this research uses RAKE. RAKE is used because in several studies it can increase accuracy [27], [32]. Then this research also used 4 splitting data, namely 60:40, 70:30, 80:20, and 90:10. Apart from that, to make it easier to carry out classification automatically, the best model from the trials was implemented using a GUI with streamlet.

B. Method

Figure 1 is the development of the model presented in the research methodology.

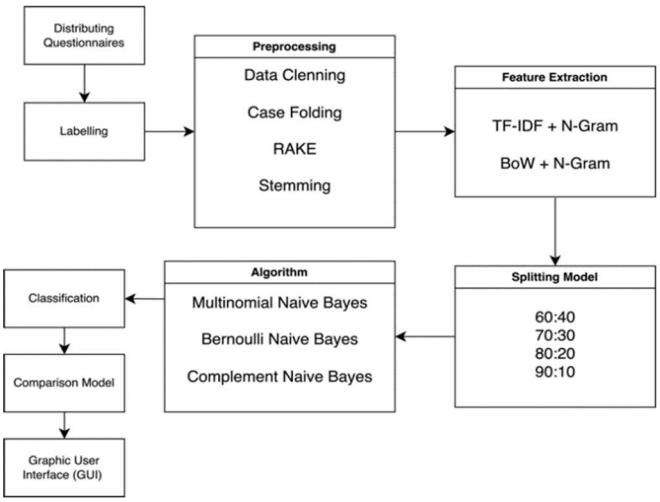

Fig. 1 Research Methodology

This study will use questionnaires distributed to teenagers aged 10–24 who are unmarried and reside in Padang. The questionnaire used to collect the dataset is presented in Table II.

TABLE III
PERSONALITY TYPE CLASSIFICATION QUESTIONNAIRE

No	Question
1	Please describe your actions when encountering new individuals in a school or campus setting and provide your rationale.
2	Please describe the type of location you frequently visit to spend your leisure time or vacation and explain your reasons.
3	Please describe your behavior when you are in a crowd or at an event and justify your reasons.
4	Please describe the type of environment you prefer for completing your college assignments and explain your reasons.
5	Please describe your approach when working on tasks that require group (team) collaboration and provide your rationale.
6	Please describe your behavior when communicating or discussing with others and explain your reasons.
7	Please describe the frequency (intensity) of your social media usage and justify your reasons.
8	Please describe your feelings when you are in a new environment (school or campus) and explain your reasons.

Based on the results of the distributed questionnaire, this study subsequently categorized the data into three labels: introvert, extrovert, and ambivert.

C. Dataset and Labelling

The dataset utilized in this study originates from questionnaires completed by respondents, who are unmarried teenagers aged 10-24 years from the city of Padang. This dataset comprises 400 samples used to investigate the personality types of teenagers. Subsequently, the dataset was labeled with three categories: Introvert, Extrovert, and Ambivert. The labeled data was then analyzed to assess its accuracy.

D. Preprocessing

This study employs several preprocessing stages, including data cleaning, case folding, RAKE, and stemming. RAKE is used to extract candidate phrases or keywords by classifying word sequences that form phrases deemed important [33]. The RAKE algorithm utilizes patterns such as the presence of conjunctions or sentence separators as potential indicators of key phrases [27]. The RAKE pseudocode is presented in Table III.

TABLE III
PSEUDOCODE RAKE

Pseudocode RAKE

```
Initialization: from rake_nltk import Rake
Input: r = Rake(language='indonesian')
    def rake_implement(text,r):
    r.extract_keywords_from_text(text)
    return r.get_ranked_phrases()
for col in X.columns:
    X[col]=X[col].apply
    (lambda text: rake_implement(text,r))
Output: X
```

The results of the pseudocode are displayed in Python software with an output dataset resulting from extracting key phrases or important terms from a text document. The results of program execution are shown in Figure 2 below:

	v1	v2	v3	v4	v5	v6	v7	v8
0	[make close friends, scold nal]	[Calm mountain atmosphere after holiday]	[active participation]	[conducive and safe]	[belongs to jobdesk]	[Feedback every time you speak]	[social media material world connect friends family]	[meeting people, nervous environment, nervous interaction]
1	[Greeting friends in a friendly neighborhood meeting establish close]	[Finished the holiday, active in the activity room]	[Very event corner looking for lots of people sitting]	[boarding house and lots of friends]	[member work]	[look for topics to talk about and build communication]	[build social connections on social media like the real world this week]	[nature walk enthusiastic enthusiastic teaching opportunities]
2	[communication with friends and be friendly]	[The natural park is open for friends to play]	[group discussion group assignments]	[I like the library as a safe place to find books]	[expert combination for group friends ideas]	[listen to friends share ideas]	[Check social media for new news at any time]	[Anxiety, anxiety arises from an environment of worry about adaptation]
3	[meet and greet people]	[It's cool after foreigners are on holiday]	[take a walk around the event venue for a road trip]	[the sound of a quiet motorbike]	[to complete the task easily]	[pay attention to people talking talking]	[social media weekends away from the screen]	[tug-of-war exploration of understanding the bargaining environment]
4	[the atmosphere of the news meeting]	[By quality time quiet and quiet busy]	[Many people come to sit in silence]	[make tasks calm, quiet, clear, focused thinking]	[help come in]	[communication listening suggestions]	[Not using social media will disrupt your productivity]	[go confused confused why]
...								
395	[positive energy group discussions like contributions]	[Popular tourist attraction, quiet, happy, delicious, happy to visit]	[Soak up moments like talking to people]	[easy to adapt to team dynamics, likes someone who understands consistency]	[likes to talk, people like to think]	[calm reflective voice authoritative speech]	[Social media and friends come in contact and the focus tends to be meaningful]	[forms of social interaction tend to be open]
396	[people are looking for interaction]	[City parks are easy to walk]	[tend to avoid paying attention to looking for groups]	[look for group work responsibilities]	[calm direction comfortable communication]	[intimate comfortable social interactions]	[social media hanging calm happy interactions]	[Anxious, often reluctant to speak]
397	[likes to talk people ideas concept]	[take part in community events for friends at home to enjoy free content festivals]	[avoids the center of attention, likes to talk in groups]	[It's easy to meet peacefully, the team doesn't like the figure]	[likes to talk to people who think deliciously]	[calm reflective voice authoritative speech]	[social media social drag]	[Sometimes you choose, sometimes it's comfortable and crowded]
398	[tend to avoid crowded social events]	[crowded and comfortable cinema]	[really think about choosing participation]	[completion of tasks productively tends to be]	[tend to avoid spilling out emotional discussions]	[tends to like oral written expression]	[Social media platforms often limit active hobby focus]	[full participation environmental process]
399	[try to create an atmosphere that supports active collaboration]	[happy to attend a private art event with friends showing off their favors]	[gathering energy looking for time to be alone is busy]	[help a team's personal vision understand active focus]	[support the ideas of people who are enthusiastic about discus]	[like hearing ideas from people who are responsive to communication]	[follow the flexible social media schedule]	[Mobile people enjoy interaction]

Fig. 2 RAKE Result

E. Feature Extraction

Feature extraction is a technique for obtaining features from a form, the values of which will later be analyzed for subsequent processes, such as classification [27]. Classification is the process of assigning an object to one of the predefined categories [28]. This study employs TF-IDF for feature extraction. TF-IDF has been frequently used by previous researchers in Natural Language Processing and Machine Learning. The development undertaken in this study involves combining TF-IDF and N-Gram for word weighting, as described by the following equation:

$$tfidf_{vectorizer} = Tfidf_{vectorizer}\left(Ngram_{range} = (2,2)\right) \quad (1)$$

$$TF - IDF + N - Gram\ Z$$

From this formula, this research produces a new equation called TF-IDF+N-Gram Z. The way the two equations are processed is processed simultaneously. This is different from previous research which was processed separately. For more details, see Table IV which is the pseudocode of TF-IDF+N-Gram Z.

TABLE IV
DEVELOPED TF-IDF+N-GRAM PSEUDOCODE

Pseudocode 8. Developed TF-IDF+N-Gram

```
Initialization: from sklearn.feature_extraction.text import TfidfVectorizer
Input: data['combined_data'] = data['v1'] + ' ' + data['v2'] + ' ' + data['v3'] + ' ' + data['v4'] + ' ' + data['v5'] + ' ' + data['v6'] + ' ' + data['v7'] + ' ' + data['v8']
tfidf_vectorizer = TfidfVectorizer(ngram_range=(2, 2))
tfidf_matrix = tfidf_vectorizer.fit_transform(data['combined_data'])
tfidf_df = pd.DataFrame(tfidf_matrix.toarray(), columns=tfidf_vectorizer.get_feature_names_out())
Output: print(tfidf_df)
```

The pseudocode above outlines the steps for implementing the TF-IDF algorithm with the use of N-grams (bigrams) utilizing the scikit-learn library on text data. The explanation is as follows:

a. Importing Required Libraries: The process begins by importing the necessary libraries, including TfidfVectorizer from scikit-learn, which provides tools to convert a collection of documents into a TF-IDF matrix. Initial preprocessing involves combining content from multiple text columns into a new column called 'combined_data'. This step ensures that text from all columns is efficiently captured and fed into the algorithm.
b. Initialization and Configuration: The next step is to initialize and configure the TfidfVectorizer object. In this case, bigrams (N-grams of length 2) are used, meaning every pair of consecutive words is considered. Using bigrams in TF-IDF provides a better understanding of the contextual relationship between words in the documents.
c. Applying TF-IDF: The TF-IDF algorithm is then applied to the combined text data using the fit_transform method. The result is a TF-IDF matrix, where each row represents a document, and each column represents the TF-IDF score for a specific bigram. This matrix is converted into a DataFrame using pandas' pd.DataFrame, including the feature names (bigram names) as column names.
d. Output the DataFrame: Finally, the resulting DataFrame is printed to provide a better understanding of the distribution of TF-IDF scores for each bigram in the documents. This output facilitates further analysis of the text representation generated by the algorithm, offering valuable insights into text processing and its linguistic characteristics.

The use of N-grams in TF-IDF offers a better understanding of the contextual relationships between words in documents. By considering word pairs, the algorithm can be more sensitive to language structure and meanings that might be overlooked when only using unigrams. In other words, N-grams allow us to capture richer contexts and more complex relationships between words in documents. This has the potential to enhance the understanding and interpretation of text analysis, particularly in tasks requiring nuanced linguistic processing, such as sentiment analysis or document clustering. Thus, the use of N-grams in TF-IDF represents an innovation that enriches the text processing approach.

F. GUI

This GUI design is designed to make it easier for users to go through the personality type classification process interactively. This application is expected to contribute to the understanding and monitoring of adolescent personality characteristics in a more efficient and user-friendly manner.

III. Results and Discussion

Figure 3 below is the result of a questionnaire that has been labeled with the three labels. From Figure 3, it can be seen that Label 1 is Introvert, Label 2 is Extrovert, and Label 3 is Ambivert, with the data distribution as follows: Introverts have 144 samples, Extroverts have 68 samples, and Ambiverts have 188 samples. After determining the quantity of each label, preprocessing was carried out using the four previously mentioned tools. Subsequently, modeling was performed using the MNB, BNB, and CNB algorithms. Below in-depth analysis for this result:

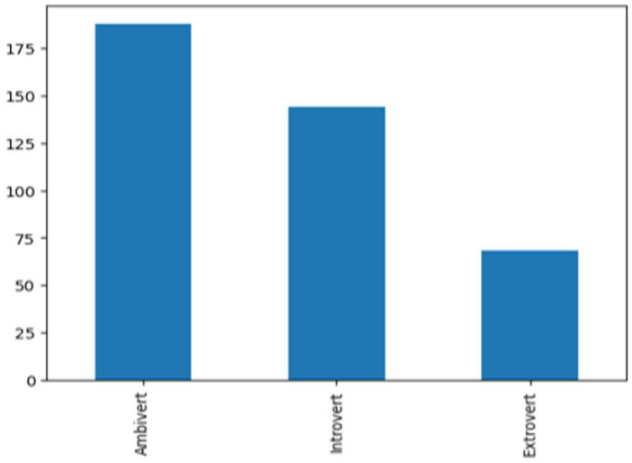

Fig. 3 Data distribution each label

1) Model Performance Comparison: The highest score indicates that the model has a higher accuracy or better predictive performance in identifying Ambivert personalities. This could suggest that the features extracted for Ambivert traits are more distinguishable and consistent, allowing the model to effectively identify individuals with this personality type. The performance for Introvert personality detection is slightly lower than Ambivert. This might indicate that the features associated with Introversion are less distinct or more variable within the dataset, making it harder for the model to classify these personalities accurately. The lowest score for Extrovert suggests that the features associated with this personality type may be the least well-defined or most similar to the other types, leading to more misclassifications or lower confidence in predictions.

2) Reasons for Better or Worse Results: The effectiveness of the machine learning model is highly dependent on the quality and relevance of the features extracted. If the features for Ambivert personalities are more distinct (e.g., consistent behavioral patterns, communication styles), the model is likely to perform better in detecting this type. Conversely, if the features for Extroverts overlap with those of other personalities or are less consistent, the model's performance will naturally be lower. If the dataset used for training is imbalanced (e.g., more data points for Ambiverts and fewer for Extroverts), the model might be biased towards the more represented class, leading to better performance in detecting Ambiverts and poorer performance for Extroverts. The model's complexity might also play a role. Simpler models might struggle to capture the nuances of personality traits, especially for more complex or overlapping categories like Extroverts and Introverts.

3) Practical Implications: In educational environments, accurately detecting Ambivert and Introvert students might help tailor learning experiences that align with their social and cognitive needs. However, the lower performance for Extroverts might require additional methods or refined features to ensure these students are not overlooked or misclassified. Understanding adolescent personality traits can be crucial for early intervention in mental health. The results

suggest that while the model is effective for certain personality types, further development is needed to ensure comprehensive and accurate detection across all personality types, especially for those with extroverted traits. To improve the model's performance, especially for detecting Extroverts, you might consider using more advanced feature extraction techniques, increasing the diversity and size of the dataset, or employing ensemble methods to capture a wider range of personality features.

In this study we did data splitting between the test data and the testing data. The data splitting that we did was 60:40, 70:30, 80:20, and 90:10. We did this to serve as a validation for the model. After training the model, we use the testing data to validate that the model is not only effective on the data it has learned but also on data it has never seen before. This validation is crucial to ensure that the model we build is not only theoretically sound but also practical and ready to be used on real-world data. By splitting the data into training and testing sets, we can be more confident that the resulting model will perform well not only on the data it has been trained on but also when faced with new data, leading to a more robust and generalizable model.

TABLE V
TF-IDF UNIGRAM ACCURACY

No	Data Split	*MNB*	*BNB*	*CNB*
1	60:40	54	74	70
2	70:30	65	77	81
3	80:20	72	83	83
4	90:10	67	80	67

Table V presents the accuracy values from TF-IDF Unigram feature extraction. Below in-depth analysis for this result in Table V:

1) Data Split Influence on Model Accuracy:

The table shows that the accuracy of each model varies depending on the data split ratio. Specifically, as the training data increases (from 60% to 90%), the accuracy of the models generally improves or remains stable, particularly for MNB and BNB. This trend indicates that these models benefit from larger training datasets, which helps them generalize better and learn more representative patterns from the data.

2) Model-Specific Performance

Multinomial Naive Bayes (MNB): ** The MNB model shows a gradual increase in accuracy as the training data increases, peaking at 80:20 with an accuracy of 72%. However, its performance declines slightly at 90:10. The MNB model is generally effective when dealing with word counts or frequency data, which may explain why its performance improves with more data but does not surpass BNB or CNB.

Bernoulli Naive Bayes (BNB): **BNB consistently outperforms the other models, achieving the highest accuracy across all data splits, with its best performance at 80:20 (83% accuracy). BNB is well-suited for binary/boolean features, which might suggest that the presence or absence of specific terms is more indicative of the personality traits being detected in your data. This model's superior performance across the board suggests that binary features are more predictive in this context than the frequency-based features used by MNB.

Complement Naive Bayes (CNB): CNB also performs well, particularly at 80:20, where it matches BNB's top accuracy of 83%. CNB is designed to address the imbalance in data and tends to work better with imbalanced datasets. Its strong performance implies that it effectively handles any class imbalance in your data, making it a robust choice for this task.

3) Comparison and Practical Implications

The BNB model's dominance suggests that the presence or absence of specific words (rather than their frequency) is crucial in detecting adolescent personality traits. In practice, this means that the binary representation of features (indicating whether a term appears or not) is more informative for this task. The strong performance of BNB across all data splits indicates that it is likely the most reliable model for this task.

Although MNB shows improvement with more data, its performance does not reach the levels of BNB or CNB. This could be due to the nature of personality detection, where the frequency of certain words does not provide as much insight as their mere presence or absence. Therefore, while MNB can still be useful, it may not be the best model for tasks that require binary decisions like personality detection.

CNB's strong performance, especially at 80:20, indicates its effectiveness in dealing with any imbalances in your data. This suggests that if your dataset is not balanced, CNB can provide more reliable results compared to MNB. In practice, CNB could be a good choice when working with datasets that may have uneven class distributions.

4) Practical Implications

The results of this study indicate that for detecting adolescent personality traits using TF-IDF unigrams, the Bernoulli Naive Bayes model is the most effective. This model's superior performance suggests that binary feature representation is more critical for this specific task. The practical implication is that when developing machine learning systems for personality detection, particularly with textual data, using BNB is likely to yield better results. Additionally, if the dataset is imbalanced, CNB should be considered as it has demonstrated robust performance under such conditions. This insight can guide future work in feature extraction and model selection, emphasizing the importance of binary feature engineering in personality detection tasks.

TABLE VI
TF-IDF BIGRAM ACCURACY

No	Data Split	*MNB*	*BNB*	*CNB*
1	60:40	75	70	70
2	70:30	74	79	74
3	80:20	69	86	79
4	90:10	98	95	98

Table VI presents the TF-IDF bigram accuracy. Below in-depth analysis for this result in Table VI. The table shows that the accuracy of the models varies significantly across different data splits. Notably, all models reach their peak accuracy at the 90:10 data split, with MNB and CNB both achieving a high of 98%, and BNB close behind at 95%. This pattern suggests that the models, especially MNB and CNB, significantly benefit from a large training dataset, allowing

them to capture more complex patterns that might be inherent in bigram features.

The Model-Specific Performance (MNB) model shows a remarkable improvement as the training data increases, particularly at the 90:10 split, where it reaches 98% accuracy. This sharp increase indicates that MNB is highly sensitive to the amount of training data when working with bigram features. The model's performance at earlier splits (60:40 to 80:20) is moderate, suggesting that MNB might struggle to generalize well with smaller datasets when bigrams are used. Bernoulli Naive Bayes (BNB): BNB demonstrates strong performance, especially at the 80:20 and 90:10 splits, where it achieves 86% and 95% accuracy, respectively. The BNB model's reliance on binary/Boolean features might be less effective with bigrams compared to unigrams, which could explain its lower performance at the 60:40 and 70:30 splits. However, as more training data is provided, the model adapts better, indicating that bigrams become more informative in larger datasets. Complement Naive Bayes (CNB): CNB performs steadily across the splits, with a notable peak at the 90:10 split where it matches MNB's accuracy of 98%. CNB's design to handle imbalanced data might not be as critical here since the results suggest that as the dataset grows, the model's ability to generalize improves, even without explicit handling of class imbalance.

The outstanding performance of MNB at the 90:10 split (98% accuracy) suggests that the model is particularly effective when it has ample data to learn from. In practical terms, this means that MNB can be a powerful tool for detecting personality traits when bigrams are used, provided that a large and diverse training set is available. The bigram features likely capture more context and subtle nuances that MNB can leverage effectively with sufficient data. While BNB doesn't outperform MNB or CNB at higher data splits, its consistent performance across splits highlights its robustness. This suggests that BNB may be a more reliable choice when the training data is limited or when the focus is on binary features. However, its slight lag behind MNB and CNB at the 90:10 split indicates that it may not fully capitalize on the contextual information bigrams provide. CNB's strong showing at the 90:10 split, matching MNB's accuracy, implies that it is highly effective with bigrams, particularly when there is sufficient data. This model's steady performance across data splits also suggests that it could be a safer choice in scenarios where data distribution is uneven, as it can handle such scenarios without a significant drop in accuracy.

TABLE VII
TF-IDF TRIGRAM ACCURACY

No	Data Split	*MNB*	*BNB*	*CNB*
1	60:40	68	65	58
2	70:30	77	77	65
3	80:20	76	86	72
4	90:10	67	93	67

The results suggest that for detecting adolescent personality traits using TF-IDF bigrams, the choice of model can be influenced heavily by the amount of training data available. If a large dataset is accessible, MNB and CNB both offer excellent accuracy, making them strong candidates for this task. However, if the dataset is smaller, BNB's robustness makes it a reliable alternative. In practical terms, this means that for tasks requiring fine-grained analysis (captured well by bigrams), MNB or CNB should be preferred when data is plentiful, while BNB remains a viable option for more constrained scenarios.

Subsequent testing was conducted using the TF-IDF and N-Gram model from previous research by Suhasini & Vimala (2021). The accuracy achieved with the previous weighting model can be seen in Table VIII.

TABLE VIII
ACCURACY WITH USING THE FOREIGN RESEARCH WORD MOVEMENT MODEL

Algorithm	Accuracy	Precision	Recall	F1-Score
MNB	93%	93%	94%	93%
BNB	96%	97%	97%	97%
CNB	95%	95%	95%	95%

It can be observed that the use of the old weighting model shows a decrease of about 2% compared to the model developed in this study. However, when compared to the research by Suhasini & Vimala, the accuracy has increased by 6%. This is due to the different data used. This study used data from questionnaires distributed among youth in Padang. The youths who filled out the questionnaires wrote more neatly, whereas the previous research used Twitter data, which often includes abbreviations or symbols that are not understood by the machine. Therefore, the questionnaire data is of higher quality compared to Twitter data.

This study also compares the results with previous researchers, as seen in Table IX. The evaluation of training and testing performance of the developed TF-IDF and N-Gram models shows better results. This demonstrates that the combination of TF-IDF and N-Gram is an effective method for developing word weighting algorithms.

TABLE IX
COMPARISON WITH PREVIOUS RESEARCH

No	Word Weighting	Algorithm	Dataset	Result
1	TF-IDF [34]	Regression, Logistic, XGBoost	16 Coordinate MB-Model	Highest accuracy 80,97%
2	TF-IDF N-Gram [22]	Naïve Bayes, SVM	Assamese textual data	Highest accuracy 80%
3	TF-IDF N-Gram [35]s	Random Forest	IMDB movie reviews	Accuracy (93.81%),
4	TF-IDF [36]	BERT	Mandarin text	Highest accuracy 90,75%
5	TF-IDF [37]	Bernoulli NB Extra-trees Classifier	Newspaper editorial containing 1604 documents	Highest accuracy 91%
6	TF-IDF [38]	Multinomial NB Extra-trees Classifier	Newspaper editorial containing 1604 documents	Highest accuracy 91%
7	TF-IDF N-Gram (This Research)	Multinomial NB Bernoulli NB Complement NB	Text of Personality	Accuracy 98% Accuracy 95% Accuracy 98%

Table IX shows the performance of the Feature Extraction model. The accuracy of the algorithms developed for feature extraction ranges between 0.95 and 0.98. The results obtained

are better than those of some previous studies. After the evaluation process, the next step is to conduct testing using a GUI. Figure 4 shows the GUI using the Multinomial Naïve Bayes model with TF-IDF Unigram feature extraction and a 90:10 data split. Below the explanation the differences between the results of this study and previous research to strengthen the novelty of this research based on Table IX.

Fig. 4 Teenage Personality Detection with GUI

Higher Accuracy: This research using the TF-IDF+N-Gram method demonstrated significantly higher accuracy rates (98%, 95%, and 98%) compared to the highest accuracy rates reported in previous studies. For example, previous studies using TF-IDF alone or in combination with N-Gram achieved accuracy levels ranging from 80% to 93.81%. This marked improvement in accuracy showcases the effectiveness of your integrated TF-IDF+N-Gram Z approach.

Algorithm Diversity: While previous studies employed a range of algorithms such as Regression, Logistic Regression, XGBoost, SVM, and Random Forest, your research specifically used Multinomial Naïve Bayes (NB), Bernoulli NB, and Complement NB. The successful application of these algorithms in your study suggests that TF-IDF+N-Gram Z can enhance performance across different classifiers, further validating the robustness of your approach.

Different Datasets: The datasets used in this research (Text of Personality) differ from those in previous studies, which included datasets like Assamese textual data, IMDB movie reviews, and newspaper editorials. Despite these differences, your method consistently outperformed prior models across various datasets, suggesting that TF-IDF+N-Gram Z is adaptable and effective across different text classification tasks.

Simultaneous Processing: One of the key novelties of this research is the simultaneous processing of TF-IDF and N-Gram, which contrasts with previous studies that treated these processes separately. This innovation likely contributed to the significant improvements in accuracy and efficiency observed in your results.

IV. Conclusion

This presents significant advancements in the field of text classification by introducing the TF-IDF+N-Gram Z method. This method, which integrates TF-IDF and N-Gram processing simultaneously, has proven to be more effective than traditional approaches where these techniques were applied separately. The integration of RAKE (Rapid Automatic Keyword Extraction) as a preprocessing tool further enhances the accuracy of the models used in this study. The main contribution of this research lies in the development of the TF-IDF+N-Gram Z method, which has resulted in a notable increase in the accuracy of personality detection models, specifically achieving up to 98% accuracy with Multinomial Naive Bayes (MNB) and Complement Naive Bayes (CNB). This method significantly outperforms previous approaches, demonstrating that the simultaneous processing of TF-IDF and N-Gram features provides a richer and more contextually accurate representation of text data in personality type detection, so that further research can be carried out to identify learning patterns or overcome juvenile delinquency and others.

Acknowledgment

We thank YPTK Padang, Indonesia for funding this research.

References

[1] F. Syuhada and R. A. Pratama, “Feature Extraction Technique For Text Mining Requirement For Reuse in Software Product Lines: A Systematic Literature Review,” SainsTech Innovation Journal, vol. 3, no. 2, pp. 87–95, Nov. 2020, doi: 10.37824/sij.v3i2.2020.230.

[2] G. R. Kumar, “A Summarization on Text Mining Techniques for Information Extracting From Applications And Issues,” Journal Of Mechanics of Continua And Mathematical Sciences, vol. spl5, no. 1, Jan. 2020, doi: 10.26782/jmcms.spl.5/2020.01.00026.

[3] H. Liang, X. Sun, Y. Sun, and Y. Gao, “Text feature extraction based on deep learning: a review,” EURASIP Journal on Wireless Communications and Networking, vol. 2017, no. 1, Dec. 2017, doi:10.1186/s13638-017-0993-1.

[4] M. K. A. Reiki, Y. Sibaroni, and E. B. Setiawan, “Comparison of Term Weighting Methods in Sentiment Analysis of the New State Capital of Indonesia with the SVM Method,” International Journal on Information and Communication Technology (IJoICT), vol. 8, no. 2, pp. 53–65, Jan. 2023, doi: 10.21108/ijoict.v8i2.681.

[5] R. Shahid et al., “Predicting Customer Sentiment in Social Media Interactions: Analyzing Amazon Help Twitter Conversations Using Machine Learning,” International Journal of Advanced Science Computing and Engineering, vol. 6, no. 2, pp. 52–56, Jul. 2024, doi:10.62527/ijasce.6.2.211.

[6] V. Vichianchai and S. Kasemvilas, “A New Term Frequency with Gaussian Technique for Text Classification and Sentiment Analysis,” *Journal of ICT Research and Applications*, vol. 15, no. 2, pp. 152–168, Oct. 2021, doi: 10.5614/itbj.ict.res.appl.2021.15.2.4.

[7] V. Talasila, M. V Mohan, and N. M. R, “Enhancing Text-to-Image Synthesis with an Improved Semi-Supervised Image Generation Model Incorporating N-Gram, Enhanced TF-IDF, and BOW

Techniques," *International Journal of Intelligent Systems and Applications in Engineering* , vol. 11, no. 7s, pp. 381–397, 2023.
[8] C. Liu, Y. Sheng, Z. Wei, and Y.-Q. Yang, "Research of Text Classification Based on Improved TF-IDF Algorithm," 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE), pp. 218–222, Aug. 2018, doi:10.1109/irce.2018.8492945.
[9] H. Fan and Y. Qin, "Research on Text Classification Based on Improved TF-IDF Algorithm," Proceedings of the 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018), 2018, doi: 10.2991/ncce-18.2018.79.
[10] B. Kabra and C. Nagar, "Convolutional Neural Network based sentiment analysis with TF-IDF based vectorization," *Pg 1 J. Integr. Sci. Technol*, vol. 11, no. 3, p. 503, 2023.
[11] M. N. Saadah, R. W. Atmagi, D. S. Rahayu, and A. Z. Arifin, Jurnal Ilmu Komputer dan Informasi, vol. 6, no. 1, p. 34, Oct. 2013, doi:10.21609/jiki.v6i1.216.
[12] F. Alzami, E. D. Udayanti, D. P. Prabowo, and R. A. Megantara, "Document Preprocessing with TF-IDF to Improve the Polarity Classification Performance of Unstructured Sentiment Analysis," Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, pp. 235–242, Aug. 2020, doi:10.22219/kinetik.v5i3.1066.
[13] I. Imelda and Arief Ramdhan Kurnianto, "Naïve Bayes and TF-IDF for Sentiment Analysis of the Covid-19 Booster Vaccine," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 7, no. 1, pp. 1–6, Jan. 2023, doi: 10.29207/resti.v7i1.4467.
[14] P. H. Prastyo, I. Ardiyanto, and R. Hidayat, "Indonesian Sentiment Analysis: An Experimental Study of Four Kernel Functions on SVM Algorithm with TF-IDF," 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), pp. 1–6, Oct. 2020, doi:10.1109/icdabi51230.2020.9325685.
[15] M. I. Alfarizi, L. Syafaah, and M. Lestandy, "Emotional Text Classification Using TF-IDF (Term Frequency-Inverse Document Frequency) And LSTM (Long Short-Term Memory)," JUITA : Jurnal Informatika, vol. 10, no. 2, p. 225, Nov. 2022, doi:10.30595/juita.v10i2.13262.
[16] Y. Pratama, A. Abdiansyah, and K. J. Miraswan, "Sentiment Analysis Using PSEUDO Nearest Neighbor and TF-IDF TEXT Vectorizer," Sriwijaya Journal of Informatics and Applications, vol. 4, no. 2, Sep. 2023, doi: 10.36706/sjia.v4i2.68.
[17] V. Suhasini and N. Vimala, "A Hybrid TF-IDF and N-Grams Based Feature Extraction Approach for Accurate Detection of Fake News on Twitter Data," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 6, pp. 5710–5723, 2021, doi:10.17762/turcomat.v12i6.10885.
[18] C. G. Jung, *The Collected Works of C. G. Jung*, vol. 5. Pantheon Books, 1956.
[19] M. C. Shehni and T. Khezrab, "Review of Literature on Learners' Personality in Language Learning: Focusing on Extrovert and Introvert Learners," Theory and Practice in Language Studies, vol. 10, no. 11, p. 1478, Nov. 2020, doi: 10.17507/tpls.1011.20.
[20] N. Rugova, "Social networks as an important part of communication in contemporary trends in adolescents, their impact on their personality and psycho-social behavior," Technium Social Sciences Journal, vol. 17, pp. 244–258, Mar. 2021, doi: 10.47577/tssj.v17i1.2873.
[21] R. Rajkumar and V. Ganapathy, "Bio-Inspiring Learning Style Chatbot Inventory Using Brain Computing Interface to Increase the Efficiency of E-Learning," IEEE Access, vol. 8, pp. 67377–67395, 2020, doi: 10.1109/access.2020.2984591.
[22] Y. Hernández, A. Martínez, H. Estrada, J. Ortiz, and C. Acevedo, "Machine Learning Approach for Personality Recognition in Spanish Texts," Applied Sciences, vol. 12, no. 6, p. 2985, Mar. 2022, doi:10.3390/app12062985.
[23] A. P. Rosyadi, W. Maharani, and P. H. Gani, "Personality Detection on Twitter User Using XGBoost Algorithm," *Jurnal Teknik Informatika (JUTIF)*, vol. 5, no. 1, pp. 69–75, 2024, doi:10.52436/1.jutif.2024.5.1.1166.
[24] I. Maliki and M. A. Sidik, "Personality Prediction System Based on Signatures Using Machine Learning," IOP Conference Series: Materials Science and Engineering, vol. 879, no. 1, p. 012068, Jul. 2020, doi: 10.1088/1757-899x/879/1/012068.
[25] M. K. Anam, T. A. Fitri, A. Agustin, L. Lusiana, M. B. Firdaus, and A. T. Nurhuda, "Sentiment Analysis for Online Learning using The Lexicon-Based Method and The Support Vector Machine Algorithm," ILKOM Jurnal Ilmiah, vol. 15, no. 2, pp. 290–302, Aug. 2023, doi:10.33096/ilkom.v15i2.1590.290-302.
[26] J. S. Baruni and Dr. J. G. R. . Sathiaseelan, "Keyphrase Extraction from Document Using RAKE and TextRank Algorithms," International Journal of Computer Science and Mobile Computing, vol. 9, no. 9, pp. 83–93, Sep. 2020, doi: 10.47760/ijcsmc.2020.v09i09.009.
[27] Z. H. Amur, Y. K. Hooi, G. M. Soomro, H. Bhanbhro, S. Karyem, and N. Sohu, "Unlocking the Potential of Keyword Extraction: The Need for Access to High-Quality Datasets," Applied Sciences, vol. 13, no. 12, p. 7228, Jun. 2023, doi: 10.3390/app13127228.
[28] C. Dev and A. Ganguly, "Sentiment Analysis of Assamese Text Reviews: Supervised Machine Learning Approach with Combined n-gram and TF-IDF Feature," *ADBU Journal of Electrical and Electronics Engineering (AJEEE)*, vol. 5, no. 2, 2023.
[29] M. Hadyan Baqi, Y. Sibaroni, and S. Suryani Prasetiyowati, "Comparative Analysis of Naive Bayes Model Performance in Hate Speech Detection in Media Social Twitter," *Jurnal Riset Komputer)*, vol. 10, no. 1, pp. 2407–389, 2023, doi: 10.30865/jurikom.v10i1.5493.
[30] S. D. Bappon and A. Iqbal, "Classification of Tourism Reviews from Bengali Texts using Multinomial Naïve Bayes," 2022 25th International Conference on Computer and Information Technology (ICCIT), pp. 270–275, Dec. 2022, doi:10.1109/iccit57492.2022.10055560.
[31] A. Kanavos, I. Karamitsos, A. Mohasseb, and V. C. Gerogiannis, "Comparative Study of Machine Learning Algorithms and Text Vectorization Methods for Fake News Detection," 2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA), pp. 1–8, Jul. 2023, doi:10.1109/iisa59645.2023.10345953.
[32] A. Bhat, C. Satish, N. D'Souza, and N. Kashyap, "Effect of Dynamic Stoplist on Keyword Prediction in RAKE," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 4, no. 6, pp. 259–264, 2018.
[33] A. M. Rukmi, D. B. Utomo, and N. I. Sholikhah, "Study of parameters of the nearest neighbour shared algorithm on clustering documents," Journal of Physics: Conference Series, vol. 974, p. 012061, Mar. 2018, doi: 10.1088/1742-6596/974/1/012061.