International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 09 Sep. 2024, pp: 75-84 www.ijaem.net ISSN: 2395-5252

Personality Test Based on Twitter (X) Posts Using Machine Learning Algorithms

Nejla Subašić[1], Ahmed Bečirević[2], Dino Kečo[3]
[1,2]*Student, International Burch University, Sarajevo*
[3]*Professor, International Burch University, Sarajevo*
Corresponding Author: Nejla Subašić

Date of Submission: 01-09-2024 Date of Acceptance: 10-09-2024

ABSTRACT

This thesis investigates the use of machine learning to predict personality traits from social media text, specifically focusing on the Myers-Briggs Type Indicator (MBTI). Using a dataset from Kaggle containing 200 tweets per user and their MBTI type, we preprocess the text data through cleaning, stopword removal, lemmatization, and handling contractions. Two vectorization techniques, TF-IDF and Word2Vec, are applied to convert the text into numerical features. Several models, including Support Vector Machines (SVM), Decision Trees, Random Forest, XGBoost, and Naive Bayes, are trained to predict individual personality dimensions (Extraversion-Introversion, Sensing-Intuition, Thinking-Feeling, Judging-Perceiving). To handle class imbalance, oversampling is employed. The models are evaluated on metrics such as accuracy, precision, recall, and F1-score, with XGBoost and Random Forest showing the best performance. A comparison of TF-IDF and Word2Vec reveals that both are effective, with varying strengths across models. An ensemble method combining Random Forest and XGBoost is also explored to improve results. This study demonstrates the potential of machine learning for personality prediction and highlights the importance of preprocessing and vectorization in achieving high accuracy.

Keywords
Myers-Briggs Type Indicator (MBTI), Personality prediction, Machine learning, Social media analysis, Natural Language Processing (NLP), TF-IDF, Word2Vec, Support Vector Machines (SVM), Random Forest, XGBoost

I. INTRODUCTION

One of the most important topics in psychological studies is personality. We can define it as a bundle of traits in people's behavior, cognition, and emotion. Personality traits are closely related to many psychological studies like identity, depression, anxiety, abuse, and poor health. One example is the strong associations found between depression and three personality traits: neuroticism, extraversion, and conscientiousness.[1] Matching a patient's personality that falls under the scope of these traits with adequate treatments will give us better results.[5]

Personality has been used to solve many practical problems in various domains. Some of those where these types of studies are most applicable are security, advertising and human resources. There has been evidence that personality constructs are strong predictors of work performance and workplace behavior. Another example is how companies try to incorporate different personality traits in one team, for mutual progress advancement and diversity. According to Psychology Today, more than 80 percent of Fortune 500 companies use some form or personality tests when hiring and training new employees. [7]

The traditional way of personality assessment so far has been questionnaires. With the internet being more and more pervasive, people are shifting towards online platforms to express themselves and interact with others. As a result, there is a growing trend of studies that focus on using users' online profiles and behaviors to predict their personalities. This way of assessing personality types has many benefits compared to the traditional method. The way people speak their minds and behave on social media provides lots of psychological content, perhaps even more that could be collected from a questionnaire. People's social media activities happen in a natural social setting and capture real interactions among friends and acquaintances, making this type of assessment less affected by experimental bias.

Another benefit is the longitudinal data social media provides, making it easier for researchers to track changes in one’s personality development. These results could be used in various practical applications such as disease prevention, online dating, targeted advertising, and personalized recommendation systems. For example, monitoring and detecting deviance of one’s social media language from his/her personality (such as a surge of expressions that reflect anxiety, depression, or suicide attempt) could help introduce early interventions that alleviate the negative impact of such deviance.[4]

A. Personality type assessment

The Myers-Briggs Type Indicator (MBTI) is a popular personality assessment tool that categorizes individuals into 16 distinct personality types based on four dichotomous scales: Extraversion (E) vs. Introversion (I), Sensing (S) vs. Intuition (N), Thinking (T) vs. Feeling (F), and Judging (J) vs. Perceiving (P). Developed by Katharine Cook Briggs and her daughter Isabel Briggs Myers, the MBTI is grounded in Carl Jung's theory of psychological types and is widely used for personal development, career counseling, and team building. Each of the 16 personality types is represented by a four-letter code (e.g., ENFP, ISTJ) that reflects an individual’s preferences in how they perceive the world and make decisions. The MBTI emphasizes that there is no "best" type, but rather that each type has its unique strengths and challenges. By understanding their MBTI type, individuals can gain deeper insights into their behaviors, motivations, and interactions with others, fostering personal growth and improving interpersonal dynamics. Despite its widespread use, the MBTI has also faced criticism regarding its scientific validity and reliability, yet it remains a highly influential tool in both psychology and popular culture. [2]

On the other hand, the Big Five personality traits—also known as the Five-Factor Model—take a dimensional approach, suggesting that personality can be described across five continuous dimensions: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (often remembered by the acronym OCEAN). The Big Five model is based on decades of psychological research and is supported by strong empirical evidence. It provides a more nuanced understanding of personality by measuring where an individual falls on each of the five dimensions, rather than categorizing them into a specific type. This approach allows for a more accurate and flexible representation of individual differences. [3]

While the MBTI offers a user-friendly, categorical perspective on personality that resonates with many for its simplicity and practical applications, the Big Five model is favored in scientific communities for its robustness, reliability, and empirical validity. Each framework has its strengths and can be useful in different contexts, but the choice between them often depends on the desired application and the level of precision needed in understanding personality traits. [3]

II. DATA

For this experiment, I am using a dataset from Kaggle that consists of 200 tweets per user, providing a comprehensive snapshot of their online behavior and language use. The dataset also includes information about each user's Myers-Briggs Type Indicator (MBTI) personality type, which categorizes them into one of the 16 personality types.[10] By leveraging this dataset, I aim to explore the relationship between language patterns in social media posts and MBTI personality types, utilizing machine learning models to predict personality traits based on textual data.

III. EXPLORATORY DATA ANALYSIS

The exploratory data analysis, as shown in the provided bar chart, reveals a significant imbalance in the distribution of personality types. The most prevalent types in the dataset are INFP and INFJ, with over 1,500 instances each, indicating a higher representation of these introspective and intuitive personality types among the users. In contrast, personality types such as ESTJ, ESFP, and ESFJ are underrepresented, with very few instances. This uneven distribution could influence the performance of machine learning models trained on this data, potentially leading to bias towards the more common types and challenges in accurately predicting the less common ones. Addressing this class imbalance is crucial for developing robust and fair models in the context of personality prediction. [11]

IV. METHOD

In this study, the focus is on training the model to distinguish between individual personality dimensions of the Myers-Briggs Type Indicator (MBTI) rather than predicting the full MBTI types. The MBTI categorizes personalities into 16 distinct types based on four dichotomous dimensions: Introversion-Extraversion (E-I), Sensing-Intuition (S-N), Thinking-Feeling (T-F), and Judging-

Perceiving (J-P). Each dimension represents a different aspect of personality, and each individual is categorized as either one or the other within each dimension. By training the model on these individual dimensions separately, several key advantages are gained, compared to predicting the full MBTI types directly.

A. Advantages of Training on Individual Dimensions

- **Improved Granularity and Specificity**: Focusing on individual dimensions allows the model to learn the unique linguistic and behavioral patterns associated with each personality trait. For instance, the language patterns of Introverts (I) and Extraverts (E) can be quite different, with Introverts potentially using more introspective and reflective language, while Extraverts might use more social and outgoing language. By training the model on these specific traits separately, it can better capture the nuanced differences in how these personality dimensions manifest in text.

- **Better Management of Class Imbalance**: When working with the full 16 MBTI types, there is often a significant class imbalance issue. Some MBTI types are much rarer than others, leading to a skewed distribution of data. This can make it challenging for the model to learn effectively, as it may become biased towards the more frequent types. By breaking down the task into four binary classification problems—one for each dimension—the class imbalance can be more effectively managed, leading to more balanced training and potentially more robust models.

- **Enhanced Interpretability**: Models trained on individual dimensions offer clearer interpretability. Instead of predicting a single MBTI type, which is a combination of four dimensions, the model can provide insights into which specific personality traits are most strongly indicated by the text data. This is valuable for applications where understanding the specific traits is more useful than just the overall type, such as in personalized marketing or tailored communication strategies.

- **Flexibility and Modularity**: Training on individual dimensions allows for more flexible and modular models. Researchers or practitioners can choose to use only certain dimensions if they are more relevant to their specific use case. For example, a study focused on decision-making styles may only be interested in the Thinking-Feeling dimension. This modular approach allows for a more targeted analysis without the need to consider the full MBTI framework.

B. Disadvantages of Predicting Full MBTI Types

- Increased Complexity and Overfitting: Predicting the full MBTI types increases the complexity of the classification problem. The model must learn to differentiate between 16 different categories, each of which is defined by a unique combination of four binary dimensions. This increased complexity can lead to overfitting, where the model becomes too tailored to the specific characteristics of the training data and performs poorly on unseen data.
- Lower Accuracy Due to Data Sparsity: With 16 different types, there is a greater likelihood of data sparsity for some types, particularly the less common ones. This sparsity can make it difficult for the model to generalize well, resulting in lower overall accuracy. In contrast, focusing on binary classifications for each dimension allows the model to have more data for each class, improving its ability to generalize.
- Reduced Interpretability: A model predicting full MBTI types provides less interpretability regarding which specific personality dimensions drive the prediction. It simply outputs a single type without explaining which dimensions contributed most to that decision. This can be a significant drawback in applications where understanding the specific traits or dimensions is crucial.
- Less Adaptability to Diverse Textual Styles: Texts written by individuals may exhibit traits from multiple dimensions in varying degrees, which are not always straightforwardly captured by a single MBTI type. By focusing on individual dimensions, the model can better adapt to the diverse textual styles and mixed signals present in the data.

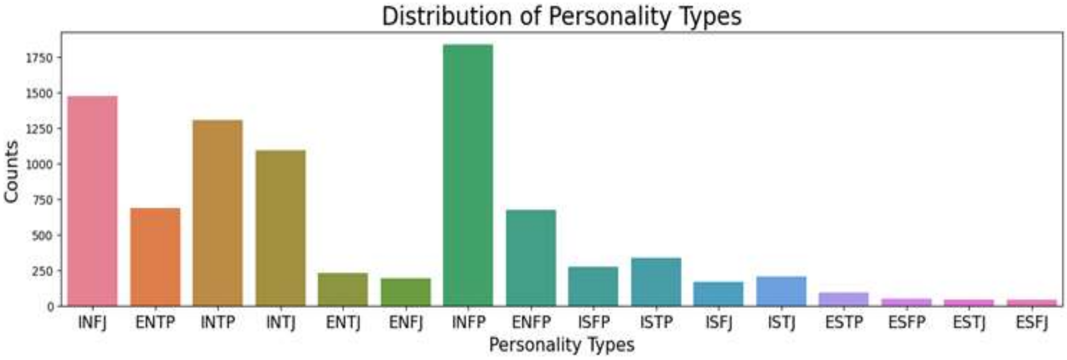

Figure 1. Distribution of Personality Types in the Dataset

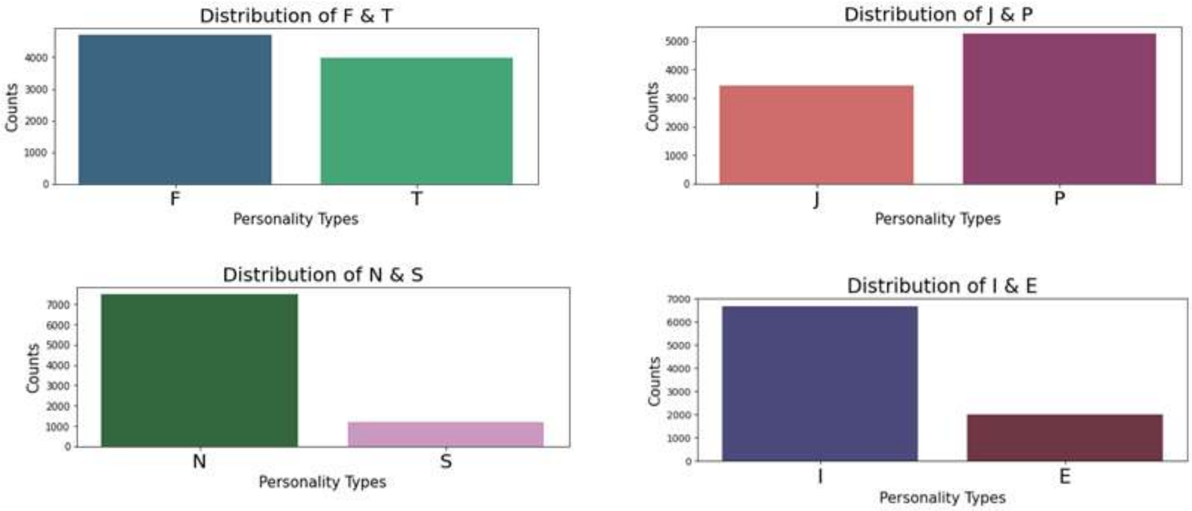

Figure 2. Distribution of individual dimensions in the Dataset

V. PREPROCESSING

In this experiment, several preprocessing steps were applied to the text data to prepare it for analysis. The preprocessing involved standardizing the text by converting it to lowercase and removing unnecessary elements such as Twitter mentions, hashtags, and URLs, which do not add value to the semantic content of the text.[12] Non-alphabetic characters were removed to focus solely on words, and extra spaces were condensed for a cleaner dataset. Additionally, contractions were expanded using the Python library ‘Contractions' to transform shortened word forms into their full equivalents (e.g., "can't" to "cannot"), ensuring more accurate text representation. [13] Short words with fewer than three characters were also eliminated to reduce noise and enhance the relevance of the remaining text data. These preprocessing steps were essential to improve the quality and consistency of the text, making it more suitable for natural language processing tasks and machine learning models.

Observing the most frequent words, bigrams, and trigrams in the dataset provides us valuable insights into the common language patterns and themes present in the users' tweets. By analyzing these frequent terms and word combinations, we can identify the topics, sentiments, and conversational styles prevalent among different personality types. [14] This information is crucial because it helps us understand the linguistic characteristics that distinguish one personality type from another, allowing for more accurate classification in our prediction model. Additionally, observing frequent n-grams (bigrams and trigrams) enables us to capture more contextually meaningful phrases, which can improve the effectiveness of our natural language processing models by retaining the syntactic and semantic nuances that single-word frequencies might miss

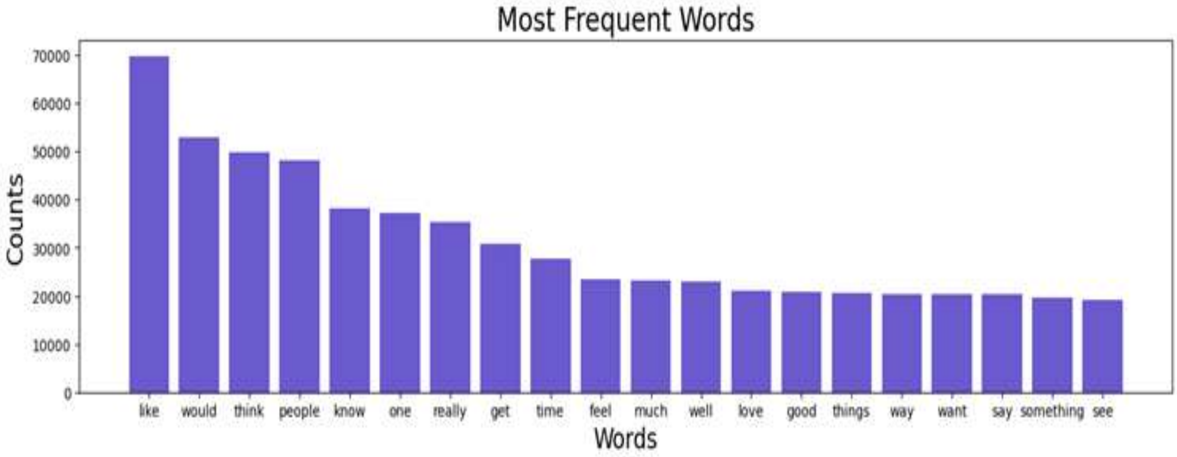

Figure 3. Most Frequent words in the user’s tweets

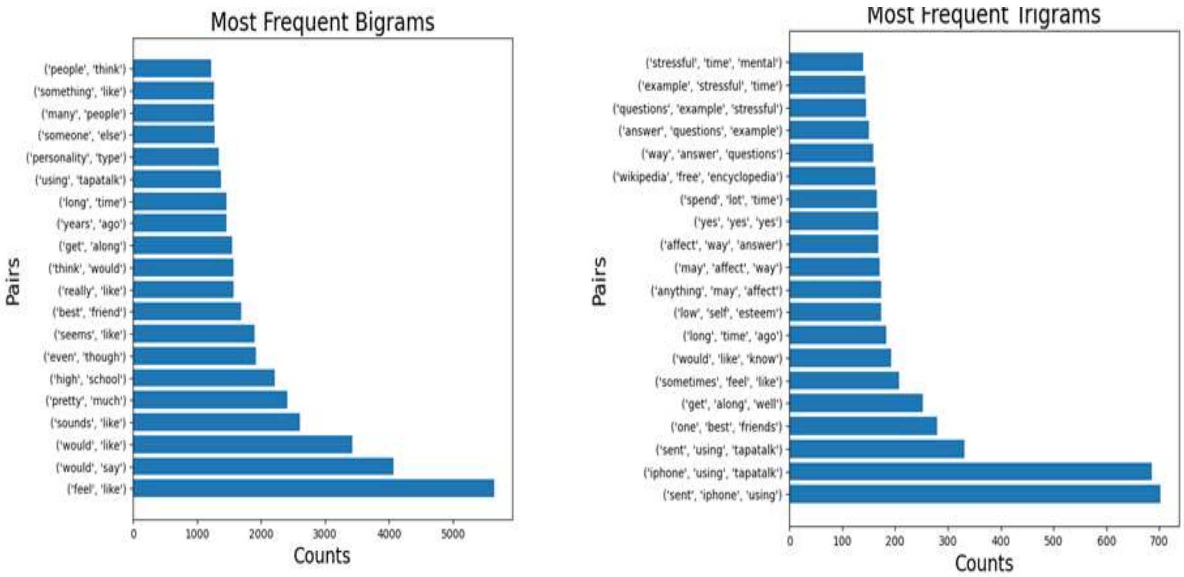

Figure 4. Most frequent Bigrams and Trigrams

Several essential text-cleaning steps were performed to prepare the dataset for building a prediction model. First, it removes stopwords using NLTK's 'word_tokenize' and a custom list of stopwords. [15] Stopwords are common words (such as "and," "the," and "is") that do not add significant meaning to text data and can be removed to reduce noise and improve model performance. After removing stopwords, the code applies lemmatization using NLTK's 'WordNetLemmatizer'. [16] Lemmatization reduces words to their base or root form, which helps in standardizing words that have different inflections but similar meanings (e.g., "running" to "run"). This process is crucial for capturing the fundamental meaning of words and ensuring that variations of the same word are not treated as different features in the model. NLTK (Natural Language Toolkit) is the primary Python library used in this code for text preprocessing, providing tools for text tokenization, lemmatization, and stopword removal. [17] These preprocessing steps help in reducing dimensionality, improving model accuracy, and enhancing computational efficiency.

VI. HANDLING IMBALANCED DATA

To address class imbalances in the dataset, oversampling was performed using the 'RandomOverSampler' from the 'imblearn' library. [18] This technique helps balance the number of samples across all classes by replicating instances from the minority classes until each class has an equal number of samples. Balancing the data is crucial for training machine learning models because an imbalanced dataset can cause the model to become biased towards the majority class, reducing its ability to accurately predict outcomes for the minority classes. By ensuring that each class is equally represented, the model is better equipped to learn from all classes, leading to improved performance and more reliable predictions

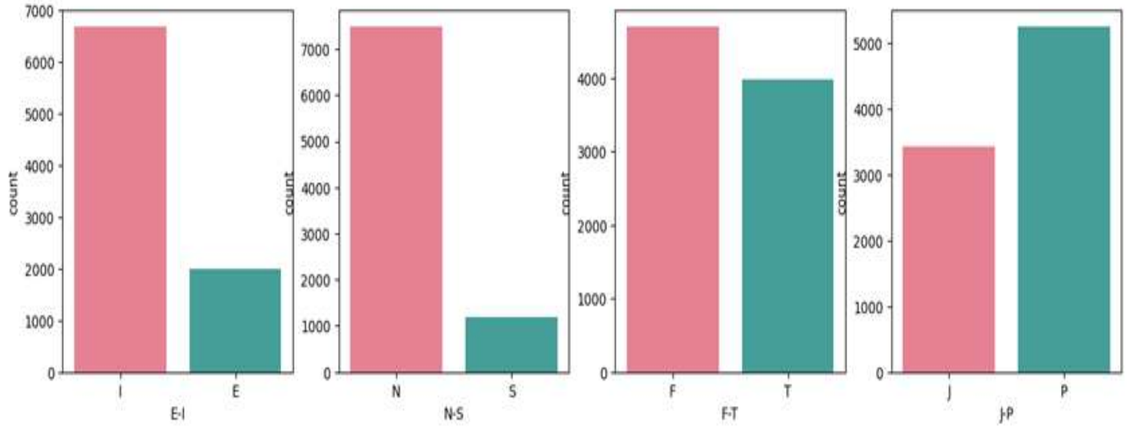

Figure 4. Distribution of dimensions before oversampling

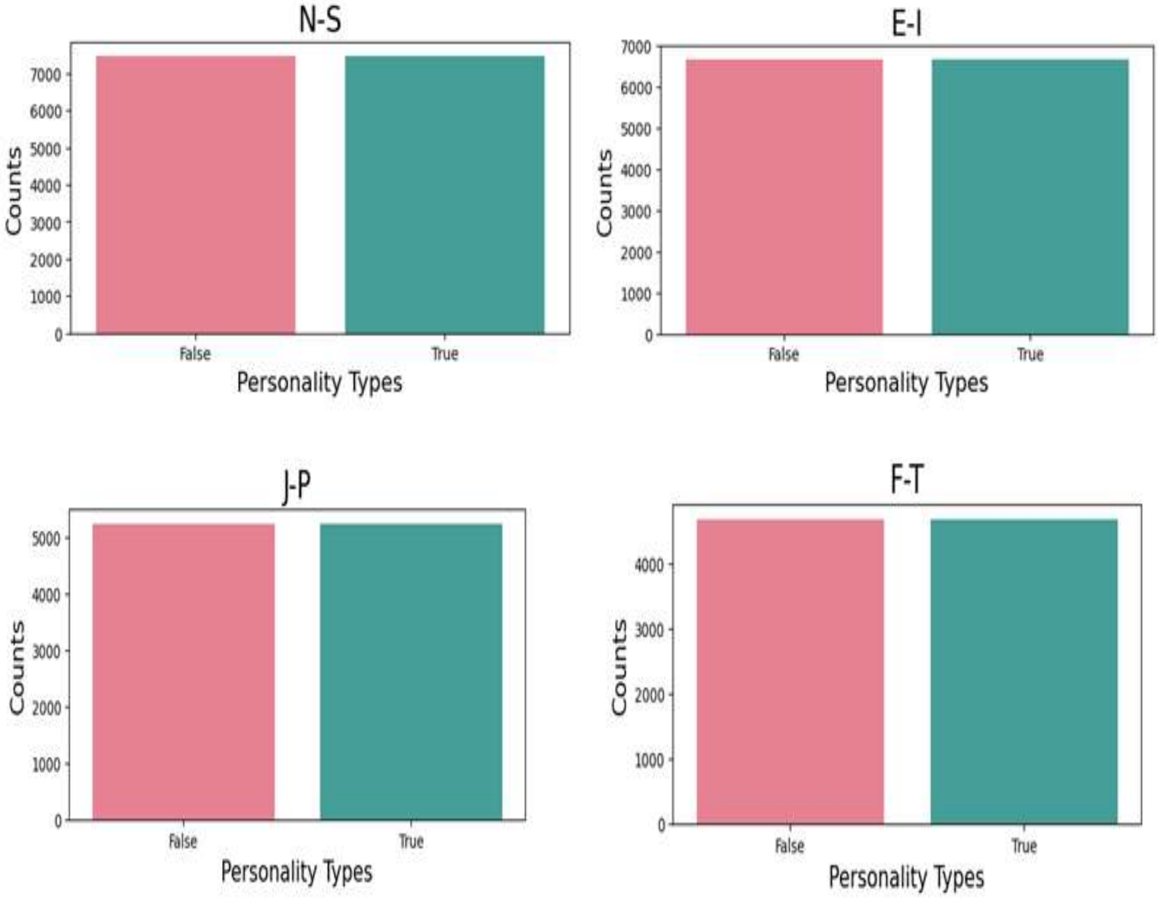

Figure 5. Distribution of dimensions after oversampling

VII. FEATURE EXTRACTION

In this experiment, two different vectorization techniques, TF-IDF (Term Frequency-Inverse Document Frequency) and Word2Vec, were employed separately to transform the text data into numerical representations for machine learning models.

TF-IDF was used to convert the tweets into a sparse matrix where each word's significance is represented by its frequency, adjusted by how unique the word is across all documents.[19] This method is effective for identifying which words are most important for distinguishing between different MBTI personality types based on their frequency of use. By focusing on term relevance, TF-IDF helps to highlight distinctive words that could be more indicative of specific personality traits.

On the other hand, Word2Vec was applied to create dense vector embeddings for each word, capturing the semantic relationships and contexts in which words appear. [20] This technique is valuable for understanding the deeper meanings behind the text, as it places similar words in close proximity within a continuous vector space, allowing the model to better grasp the nuances and subtleties of language that might correlate with different personality types.

By separately utilizing TF-IDF and Word2Vec, the goal was to compare the performance of these two vectorization methods to determine which one provides better results for the task of personality prediction. This comparison allows for a better understanding of which text representation method is more effective in capturing the linguistic patterns associated with different MBTI types.

VIII. RESULTS

For the prediction model, several machine learning algorithms were utilized to classify MBTI personality types based on tweet data. The models included a Support Vector Machine (SVM) with a linear kernel to handle high-dimensional feature spaces, a Decision Tree classifier to identify patterns through decision-making rules, a Random Forest classifier with 750 estimators to enhance prediction accuracy through an ensemble of decision trees, and an XGBoost classifier known for its speed and performance in handling large datasets and complex patterns. These diverse models were chosen to explore various approaches to text classification and to determine the most effective model for predicting personality types.

After applying Word2Vec for vectorization, training a model using Naive Bayes wasn’t possible because Naive Bayes requires non-negative feature values, such as word frequencies or TF-IDF scores. However, Word2Vec generates continuous, dense vector representations of words, which can include negative values. This violates the assumptions of the Naive Bayes algorithm, making it unsuitable for use with Word2Vec representations. Therefore, the Naive Bayes model was omitted from this analysis.

The results using the Word2Vec vectorizer show varied performance across different models and MBTI dimensions. XGBoost and Random Forest consistently outperform other models, particularly in the 'E-I' and 'N-S' dimensions, with XGBoost achieving the highest accuracy, precision, recall, F1-score, and ROC-AUC scores in these categories. This indicates that XGBoost and Random Forest are better at capturing the patterns in the Word2Vec-encoded data for these dimensions. SVM shows moderate performance overall, performing best in the 'F-T' dimension but lagging behind in others. The Decision Tree classifier exhibits the lowest performance across most metrics, particularly in the 'J-P' dimension, where it shows the weakest results. These findings suggest that more complex ensemble models like XGBoost and Random Forest benefit more from Word2Vec embeddings, likely due to their ability to handle non-linear relationships in the data.

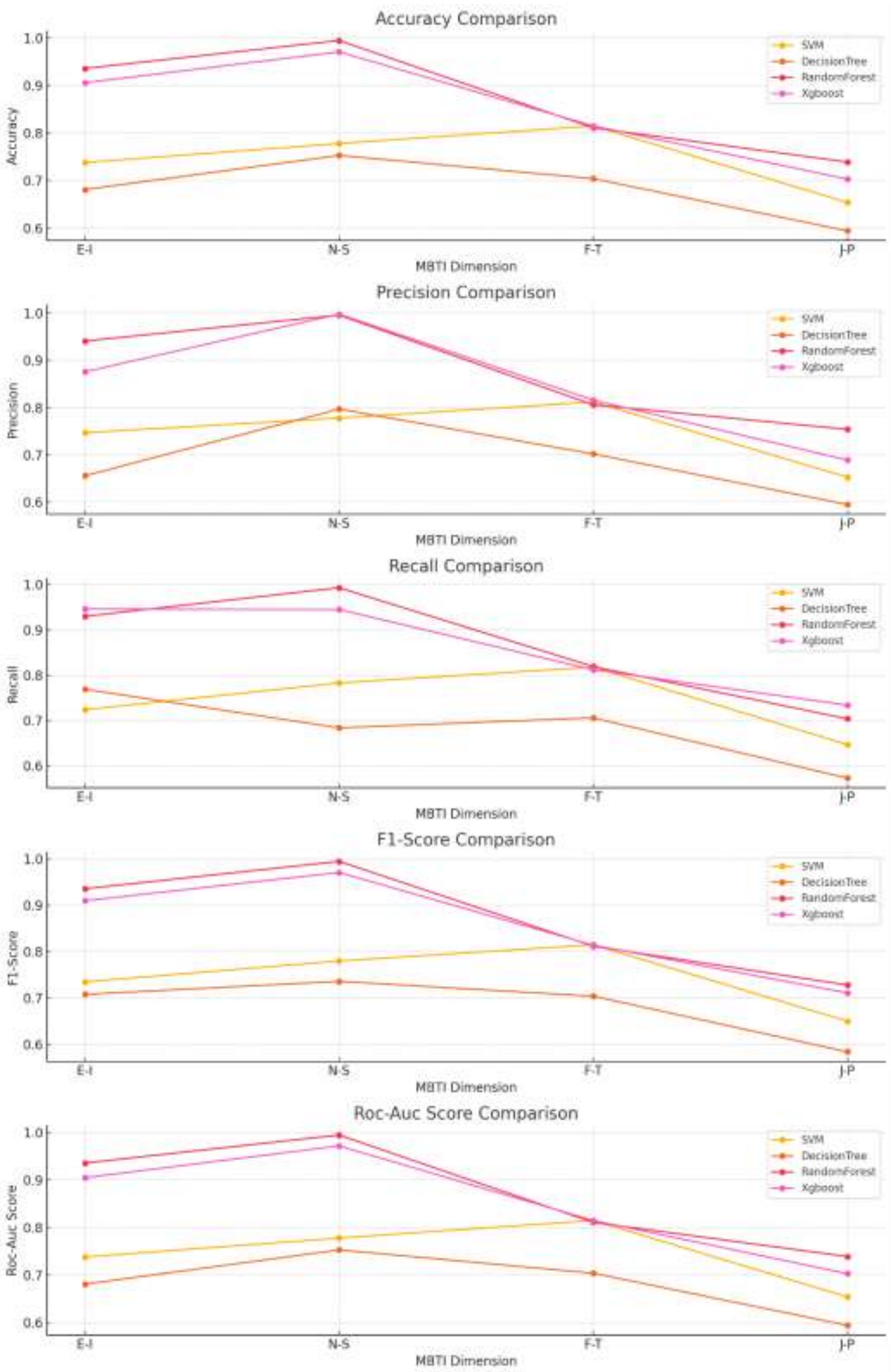

Figure 6. Results of the five trained models with Word2Vec vectorizer

The results for the TF-IDF vectorizer indicate strong performance across different models and metrics for predicting MBTI dimensions. For the E-I dimension, Random Forest and XGBoost show the highest accuracy (0.953 and 0.942, respectively) and high F1-scores (0.951 and 0.944). This suggests these models effectively balance precision and recall. For the N-S dimension, Random Forest also performs exceptionally well with an accuracy of 0.994 and an F1-score of 0.994, closely followed by XGBoost. For the F-T dimension, SVM shows competitive performance with an accuracy of 0.855 and an F1-score of 0.856, slightly outperforming other models. However, for the J-P dimension, the performance of all models drops, with Random Forest and XGBoost again leading but only achieving F1-scores around 0.818 and 0.846. Notably, Naive Bayes performs relatively well on the N-S dimension (accuracy of 0.904), but its overall performance across other dimensions and metrics is lower compared to Random Forest, XGBoost, and SVM. This analysis suggests that while Random Forest and XGBoost are the most robust models for most MBTI dimensions, SVM also shows good potential, particularly in situations where a simpler model might be preferred.

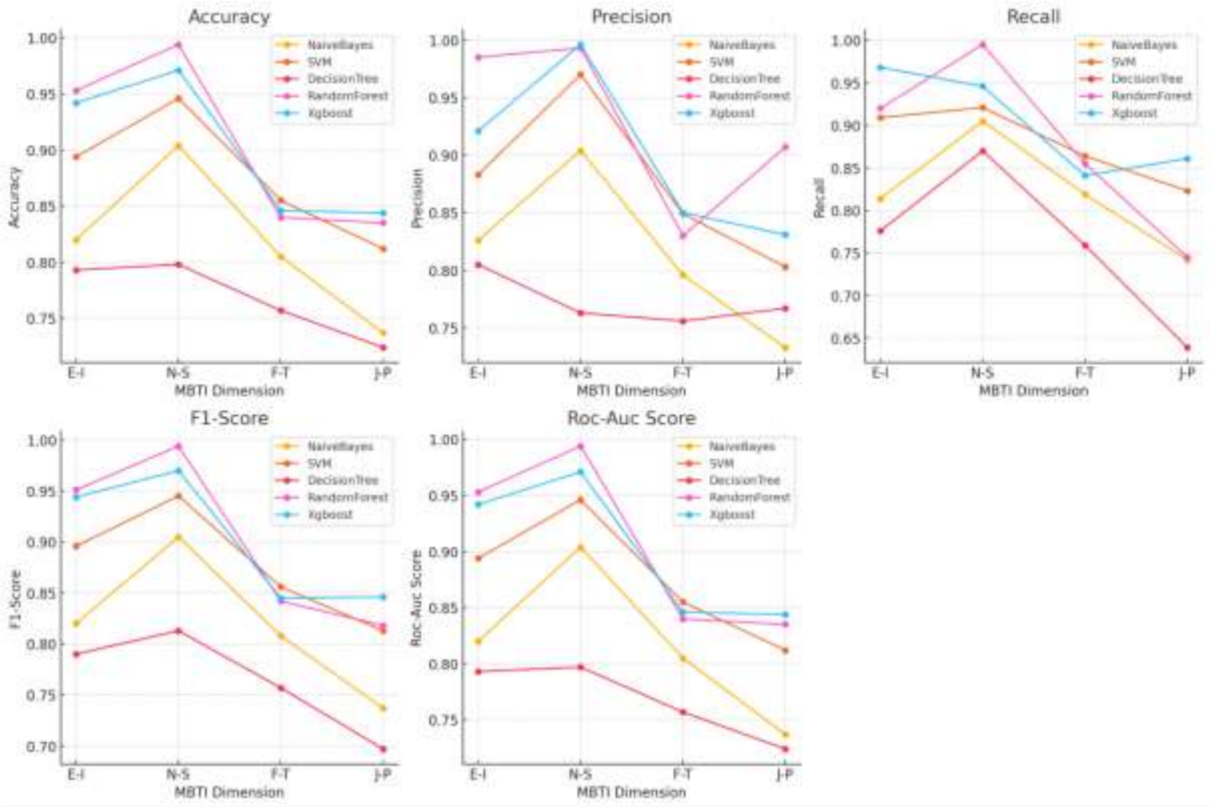

Figure 7. Results of the five trained models with TF-IDF vectorizer

IX. CONCLUSION

Based on the results from both TF-IDF and Word2Vec vectorization techniques, it is evident that the choice of vectorization has a significant impact on the performance of the machine learning models used for predicting MBTI personality types. The TF-IDF vectorizer consistently yielded higher performance metrics across most models and personality dimensions, particularly with the Random Forest and XGBoost models, which demonstrated superior accuracy, precision, recall, and F1-scores. In contrast, the Word2Vec vectorizer showed comparatively lower performance, although the Random Forest and XGBoost models still achieved relatively good results. This suggests that TF-IDF's ability to capture term frequency and inverse document frequency provided more useful features for this text classification task compared to the dense vector representations produced by Word2Vec.

Overall, the findings indicate that while both vectorization methods have their merits, TF-IDF, in combination with ensemble methods like Random Forest and XGBoost, was more effective for the MBTI personality prediction task in this dataset. Future work could explore the integration of more sophisticated embedding techniques, such as BERT or GPT-based embeddings, to potentially enhance model performance further.

REFERENCES

[1]. Al Hanai, T., Ghassemi, M., & Glass, J. (2018a, September 2). Detecting depression with audio/text sequence modeling of interviews. Interspeech 2018. Interspeech 2018. https://doi.org/10.21437/interspeech.2018-2522

[2]. Pittenger, D. J. (1993). The utility of the Myers-Briggs Type Indicator. Review of Educational Research, 63(4), 467–488.

[3]. Celli, F., & Lepri, B. (n.d.). Is Big Five better than MBTI? A personality computing challenge using Twitter data. Retrieved December 20, 2023, from https://ceur-ws.org/Vol-2253/paper04.pdf

[4]. Chandran, D., Robbins, D. A., Chang, C.-K., Shetty, H., Sanyal, J., Downs, J., Fok, M., Ball, M., Jackson, R., Stewart, R., Cohen, H., Vermeulen, J. M., Schirmbeck, F., de Haan, L., & Hayes, R. (2019). Use of Natural Language Processing to identify Obsessive Compulsive Symptoms in patients with schizophrenia, schizoaffective disorder or bipolar disorder. Scientific Reports, 9(1), 14146.

[5]. Corcoran, C. M., & Cecchi, G. A. (2020). Using Language Processing and Speech Analysis for the Identification of Psychosis and Other Disorders. Biological Psychiatry. Cognitive Neuroscience and Neuroimaging, 5(8), 770–779.

[6]. Jang, J., Yoon, S., Son, G., Kang, M., Choeh, J. Y., & Choi, K.-H. (2022). Predicting Personality and Psychological Distress Using Natural Language Processing: A Study Protocol. Frontiers in Psychology, 13, 865541.

[7]. Jayaratne, M., & Jayatilleke, B. (2020). Predicting Personality Using Answers to Open-Ended Interview Questions. IEEE Access, PP(99), 1–1.

[8]. Li, W. (2021). Predicting MBTI personality type of Twitter users [Rutgers University-Camden Graduate School]. https://doi.org/10.7282/t3-75wc-2x18

[9]. Pavan Kumar, K. N., & Gavrilova, M. L. (2019). Personality Traits Classification on Twitter. 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 1–8.

[10]. Mitchell, J. (2017). (MBTI) Myers-Briggs personality type dataset [Data set]. https://www.kaggle.com/datasets/datasnaek/mbti-type

[11]. Spelmen, V. S., & Porkodi, R. (2018). A review on handling imbalanced data. 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), 1–11.

[12]. Krouska, A., Troussas, C., & Virvou, M. (2016). The effect of preprocessing techniques on Twitter sentiment analysis. 2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA), 1–5.

[13]. Contractions. (n.d.). PyPI. Retrieved September 1, 2024, from https://pypi.org/project/contractions/

[14]. Jain, A. (2024, February 5). N-grams in NLP - Abhishek Jain. Medium. https://medium.com/@abhishekjainindore24/n-grams-in-nlp-a7c05c1aff12

[15]. NLTK :: nltk.tokenize package. (n.d.). Retrieved September 1, 2024, from https://www.nltk.org/api/nltk.tokenize.html

[16]. NLTK :: nltk.stem.WordNetLemmatizer. (n.d.). Retrieved September 1, 2024, from https://www.nltk.org/api/nltk.stem.WordNetLemmatizer.html?highlight=wordnet

[17]. NLTK :: Natural Language Toolkit. (n.d.). Retrieved September 1, 2024, from https://www.nltk.org/

[18]. RandomOverSampler — version 0.12.3. (n.d.). Retrieved September 1, 2024, from https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html

[19]. TfidfVectorizer. (n.d.). Scikit-Learn. Retrieved September 1, 2024, from https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[20]. Word2vec. (n.d.). TensorFlow. Retrieved September 1, 2024, from https://www.tensorflow.org/text/tutorials/word2vec