CROSS-PLATFORM PERFORMANCE OF MACHINE LEARNING MODELS ON REDDIT AND TWITTER

ORCUN ERDEM

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

876525

COMMITTEE

dr. Chris Emmery
dr. Dimitar Shterionov

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24th, 2024

WORD COUNT

8155

ACKNOWLEDGMENTS

I would like to thank my supervisor dr. Chris Emmery for his help and feedback during the creation of my thesis. He provided me with helpful feedback and information for me to be able to continue working on my research. I would also like to thank Tilburg University for this opportunity.

CROSS-PLATFORM PERFORMANCE OF MACHINE LEARNING MODELS ON REDDIT AND TWITTER

ORCUN ERDEM

Abstract

This thesis investigates the use of multiple different machine learning models to predict attributes of an author on social media platforms. The attributes are gender and the MBTI personality type. Data from two different social media platforms will be used. Firstly Reddit, which is a fairly lesser researched social media platform that has been used for author profiling tasks. Secondly Twitter, which is a more widely used social media platform and has also been used for multiple PAN tasks. These profiling tasks will firstly only focus on both domains individually, by predicting the attributes of an author with using the data as training and testing. Afterwards the cross-domain performance of the models will be evaluated, to see how they perform by using one platform as training data and the other as testing. Four different models will be made and tested to predict the attributes of the user. These are Logistic Regression, Random Forest, Linear SVC and BERT. Before using the data in combination with these models, some preprocessing steps will be taken. So for example URLs and retweets are removed from the Twitter dataset. Methods used in combination with the models are Term Frequency-Inverse Document Frequency (TF-IDF) in combination with word and character n-grams to transform the textual data in numerical vectors, which can be used in the models. To find the optimal parameters of a model, GridSearch has been ran with some possible parameters in combination with a cross-validation of 3, to increase the validity of the parameters. The results show a highest accuracy of 0.90 and 0.76 for gender on Reddit and Twitter, respectively. With the highest accuracy on a single dimension in the MBTI indicator of 0.95 and 0.80 on Reddit and Twitter respectively. The cross-domain accuracy of models is decreased compared to single domain performance. Which indicates models do not generalize as well cross-domain.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

1.1 *Source/Code/Ethics/Technology Statement Example*

Data Source:

The Reddit data has been acquired from dr. Chris Emmery by submitting a data agreement. Twitter data has been acquired by the help of dr. Chris Emmery, but is from TwiSty. A corpus to aid for research in author profiling, by Ben Verhoeven, Walter Daelemans and Barbara Plank. Work on this thesis did not require collecting data from human participants or animals. Consent for Reddit has been given, after signing a data agreement, Twitter data was available. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. The thesis code can be accessed through the GitHub repository following the link ['https://github.com/NucroQ/Author-Profiling']. Some code has been taken from ['https://github.com/google-research/bert'], but rewritten in other ways. All figures belong to the author of this thesis. Code has been written by the author, with the help of online documentation.

2 INTRODUCTION

In today's digital age, social media has millions of users across a variety of different platforms. This creates a great amount of data which can be interesting for author profiling tasks. Author profiling involves analyzing textual data to gain information about the characteristics of the author. These characteristics could be age, gender, nationality, political preference or personality types. By using natural language processing (NLP) techniques and machine learning algorithms, it is possible to extract characteristics of the author (Emmery et al., 2017). Because social media is becoming more popular and used by the day, it makes it possible to profile these users on social media.

While author profiling has been used for many tasks in NLP, like gender classification, there is also a growing interest in classifying personality traits of authors (Mirkin et al., 2015). One of the well-known typologies is the Myers-Briggs Type Indicator(Briggs Myers & Myers, 2010). The Myers-Briggs Type Indicator consists of four different dimensions which form the personality of a person.

Although author profiling on social media platforms is an active line of research, these studies are mainly done on a singular domain. By using two different datasets, one from Reddit and one from Twitter, this thesis

aims to provide more insights into the usability of models for cross-domain usability. The following research questions are formulated for this research:

> *How do machine learning models(LR, RF, Linear SVC and BERT) trained on Twitter data perform on Reddit data for author profiling tasks, with the focus on gender and MBTI personality traits?*

SUB-RQ1 *Which of the models(LR, RF, Linear SVC and BERT) are most accurate on only Reddit or Twitter data?*

SUB-RQ2 *Does the cross-domain performance of models change if they are trained on Reddit or on Twitter data?*

This thesis addresses two significant gaps in the current research. Firstly, the relatively lesser studied social media platform: Reddit. While Twitter and Facebook are more extensively studied for author profiling tasks, Reddit is only recently becoming more popular for these tasks. The Reddit dataset used in this study will be new. The dataset is from Emmery et al. (2024) which is recently created and unexplored. By including this new Reddit dataset, this thesis aims to broaden the scope of author profiling research across different social media datasets.

Additionally, this thesis explores the cross-domain usability of machine learning models. Given the limited amount of literature on the generalizability of machine learning models trained on one social media dataset to predict on another, this study aims to determine whether models can be effectively used for such tasks. This can be particularly useful when data from one social media platform is scarce. By training models on Reddit data and evaluating their performance on Twitter data, and the other way around, this study aims to contribute to the cross-domain usability of different machine learning models for author profiling tasks.

2.1 *Relevance*

This research is relevant from both scientific and societal perspectives. Looking at societal relevance, identifying the attributes of an author can play an important role for businesses (Plank & Hovy, 2015). Being able to accurately predict an user's gender and personality traits, businesses can accurately target their marketing campaigns to social media users. Although this could lead to ethical concerns regarding user privacy. Another benefit is for personalized content recommendations based on the user's profile, which can lead to better customer engagement and satisfaction. Lastly, Preoţiuc-Pietro et al. (2015) have conducted research on the connection between personality types and psychological disorders of social

media users. Being able to accurately predict the personality of an user could help addressing this issue. From a scientific perspective, this study contributes to the cross-domain usability of machine learning models for author profiling tasks. This provides insights into the generalizability of models across different domains, which can be beneficial to predict on a social platform where data is scarce or costly to obtain.

3 RELATED WORK

Author profiling is the task of predicting certain characteristics of authors. The characteristics can be gender, age, nationality, education or personality traits (Reddy et al., 2017). The focus in this thesis will lie on gender and personality trait predictions. The personality trait predictions are based on the Myers-Briggs Type Indicator (MBTI) (Briggs Myers & Myers, 2010). The MBTI indicator consists of four different dimensions, with each dimension having two different labels. This means totally there are 16 different personality types with the MBTI personality type indicator.

3.1 *Gender classification*

One of the main tasks in author profiling is gender prediction. This has been done in many different studies and has increased in accuracy over time. In the 2018 gender classification task of PAN the best result was 0.82 (Rangel et al., 2018). This task was from a multimodal perspective, so the Twitter corpus used in this PAN task included both textual information, but the participants were also provided with images. While the corpus also contained multiple different languages, namely English, Spanish and Arabic. The best result on only textual data was achieved by using combinations of different n-grams and the use of SVM and Logistic Regression. Most participants of this task removed URLs, usernames and hashtags. With some also lowercasing all the words. The study of Vashisth and Meehan (2020) compared different machine learning models and feature generation techniques. Models being used were, LR, MLP, SVM, NB, RF and XGBoost. While the different feature extraction techniques were TF-IDF, Word2Vec and GloVe. The higest accuracy they achieved was 57.14%. This seems low comparing to other literature and previous PAN tasks. Although it is said this might be due to the database being used in their study.

A way to achieve different results can be to focus on the feature extraction. Ameer et al. (2019) have extracted features by the use of traditional n-grams and syntatic n-grams of POS, aswell as a combination of words and characters. Because the data consists of words and words are multiple characters, the arrangements of words and characters can keep important

information about the author (Ameer et al., 2019). The highest accuracy was achieved by the use of this feature extraction technique.

The goal of the study done by Lain and Zalzala (2023) was to find a model which was trained on data from one domain and then accurately predicts characteristics of the author on another domain. Firstly models are trained and tested on a single domain and afterwards models are tested on two different domains which are indepentent from the training data. The models used were SVM, FFNN, CNN and XLNet. XLNet is a pre-trained transfer learning model. For preprocessing, URLs are removed, aswell as hashtags and numbers. Afterwards all words are transformed to lowercase lettering. Different feature extraction methods have been used in combination with the models, namely TF-IDF and Word2Vec. SVM was consistent across all different twitter datasets, but XLNet performed best on both single domain, and cross-domain data. The SVM model used was a Linear SVC. The study done by Dias and Paraboni (2020) was focused gender identification of the author in the Brazilian Portuguese language. On a single domain, the Logistic Regression in combination with TF-IDF was the best performing model. There was also a test for using different training data than test data. By using E-gov data as training data and Facebook as test data, the F1 loss was 0.01 compared to Facebook data only.

The research done by Verhoeven et al. (2016) is about classification task on a Twitter dataset, TwiSty. The corpus consists of multiple different languages including German, Italian, Dutch, French, Portuguese and Spanish. To create a model predicting the gender of the author on Twitter, they made use of a Linear SVC model. This model was implemented with standard parameters, as GridSearch on the C parameter did not improve their results. URLs and usernames are normalized to placeholders with binary features for word n-grams and character n-grams. They performed a 10-fold cross-validation to evaluate the model. The F-scores across all languages ranges from 73, till 87. With the Spanish language having the highest F-score of 87.62, and Italian being the lowest with a F-score 73.29.

One study that has been done by Alzahrani and Jololian (2021) was focused gender prediction of an author using a pretrained model. The model used was BERT, a transfer learning model. This does not require much focus on feature engineering and is mainly focused on the way you preprocess the data. Five different methods of preprocessing have been applied by them, followed by the exact same BERT model with the same hyperparameters; ran on 3 epochs, a batch size of 32, max text length of 100 and a learning rate of 2e-5. Firstly a preprocessing method is to apply no preprocessing steps. Then the steps range from only removing mentions, retweets en hashtags, to also remove URLs,

punctuations, and stopwords. The interesting outcome is that the model with no preprocessing, so handling the text as it is, is achieving the highest accuracy. With accuracy decreasing for every extra step of preprocessing. A possible explaination for this could be due to it being a pretrained model and that those perform better on larger texts and need every word and token to learn.

3.2 *MBTI prediction*

In author profiling tasks, gender and age are the most commonly predicted characteristics of the author. However, there are also others attributes that can be predicted. One of these is personality. Personality can be determined by MBTI personality types. The Myers-Briggs Type Indicator is a personality indicator with four different dimensions. Each dimensions consists of two different labels, totalling 16 different personality types. The interest in personality prediction from social media is increasing (Nguyen et al., 2016). Being able to predict an author's personality on social media can give opportunities for targeted social media marketing or even to diagnose psychological disorder (Preoţiuc-Pietro et al., 2015).

Gjurković and Šnajder (2018) predicted the personality types of Reddit users according to their MBTI types. They used a dataset consisting of a subreddit where users are able to use flairs to indicate their MBTI type. They used LIWC for the linguistic features for feature extraction. To predict the MBTI type of the author, they solve four binary classifications problems. MBTI consists of four different dimensions, with each having two different choices. Namely; Introversion/Extraversion, Sensing/iNtuition, Thinking/Feeling and Judging/Perceiving (Briggs Myers & Myers, 2010). The four independent classification problems are then combined together to get a full MBTI prediction. Different models used in their study are SVM, LR and MLP. In 82% of the cases they are able to predict the correct MBTI type, or have 1 dimension false. Two or more dimensions have been predicted correctly in 97% of the cases.

There has also been a study on personality prediction with the MBTI indicator done by Nisha et al. (2022). They used a dataset consisting of tweets from users. Every dimension from the MBTI indicator was predicted individually. Models used to predict the types were Naïve Bayes, SVM and XGBoost. With Naïve bayes performing worst and XGBoost being the best performing model, with the highest accuracy of 90% on S/N dimension and the lowest accuracy of 80% on J/P dimension. It is clear from the results that the J/P dimension has the lowest accuracy across all models used.

Research by Plank and Hovy (2015) is about MBTI prediction based on Twitter data. It is a database of 1.2M Tweets, where they will create a model to predict the MBTI personality type of the author. For the features in their model, they used binary word n-grams. URLs, hashtags and usernames are replaced with unique tokens, as pre-processing steps. Their models performs better than a majority baseline model for the I/E and T/F dimensions. With the baseline model and their model both scoring high on the S/N dimension, but with minimal difference.

While the research by Verhoeven et al. (2016) was also on gender classification, they also predicted the MBTI personality type of the authors on Twitter. The same Linear SVC model was used to predict the personality type. It was split up in four different dimensions, which has been done in the same way in other previous literature. It seems gender classification worked well, however the personality trait identification was more difficult. Their results are in compliance with other research that the E/I and T/F dimensions are easier to predict from social media.

Keh, Cheng, et al. (2019) conducted research by using BERT to predict the MBTI personality type of an author. He used posts from a online forum, where all 16 different MBTI personality types are placed on different sections of the forum so no labelling would be needed. They analyzed the data and over 95% of the users who post in a certain section identify as someone with that respective MBTI type. As preprocessing steps symbols that are not numbers, punctuations or letters have been removed and everything has been converted to lowercase letters. BERT's custom tokenizer has been used which masks and pads the sentences. The BERT model used in the research is the 'bert-base-uncased' model. There have been experiments with three different parameters, namely learning rate, max sequence length and epochs. Batch size has been kept constant at 32. The highest accuracy was achieved by using a max sequence length of 128, learning rate of 10e-5 and 30 epochs. According to them, increasing epochs had a bigger impact than changing learning rate, which only resulted in a difference in accuracy of 0.01. Results also indicate that E/I and T/F are more easily predicted by the BERT model compared to the other dimensions.

Gjurković et al. (2021) have conducted a study on personality prediction on a new dataset named 'PANDORA'. The dataset consists of comments posted by over 10,000 users. These users are annotated with personality traits and demographic information. The personality traits are labeled according to the Big 5, MBTI and Enneagram types. The Big 5 is another personality indicator which consists of the following traits; Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism (McCrae & John, 1992). The models used in the research consist of LR, SVM, RF,

NN and BERT. With BERT being the top performing model for personality prediction tasks.

Another study done on MBTI prediction is from dos Santos and Paraboni (2022). Their goal was to predict personality based from textual data based on two different social media datasets. Firstly on a Reddit datasets and secondly on a Twitter dataset. The only preprocessing of the text was the removal of special characters, the rest of the text was left unchanged. Two different models are tested, firstly BERT, which is the main focus of this research. BERT is fine-tuned by them to optimize it for MBTI classification. The second model tested is LSTM. Results show that BERT has an F1 score ranging from 0.94 to 0.89 for the MBTI predicition on Reddit data. Results are similar for the Twitter dataset, but ranging lower on certain languages, since the Twitter dataset contained different languages.

3.3 *Conclusion of works*

After evaluating all the previous literature on both gender and personality type classification tasks on social media, it is clear that Reddit is a lesser used social media platform compared to other social media data. But the main gap in the literature this thesis will address be the cross-platform application of machine learning models.

After reviewing past research, the following models have been chosen to answer the research questions. Firstly Logistic Regression, since it is widely used in previous literature and easy to interpret and understanding with decent performance. Secondly Random Forest, which is a robust machine learning model and fairly minimally used in previous literature. Linear SVC has also been chosen as one of the models. This is a SVM, but less computationally expensive, which shows good performance in the literature. Finally BERT, which is a newer pre-trained model. Literature shows BERT is performing good on personality prediction tasks and gender classification tasks. More details about the models will be explained in section 4.3.

4 METHOD

This section will highlight and elaborate on the methodology used to answer the research questions. The datasets, preprocessing steps, models and evaluation metrics will be explained and visualized. Figure 1 displays a flowchart of the different parts of the methodology.

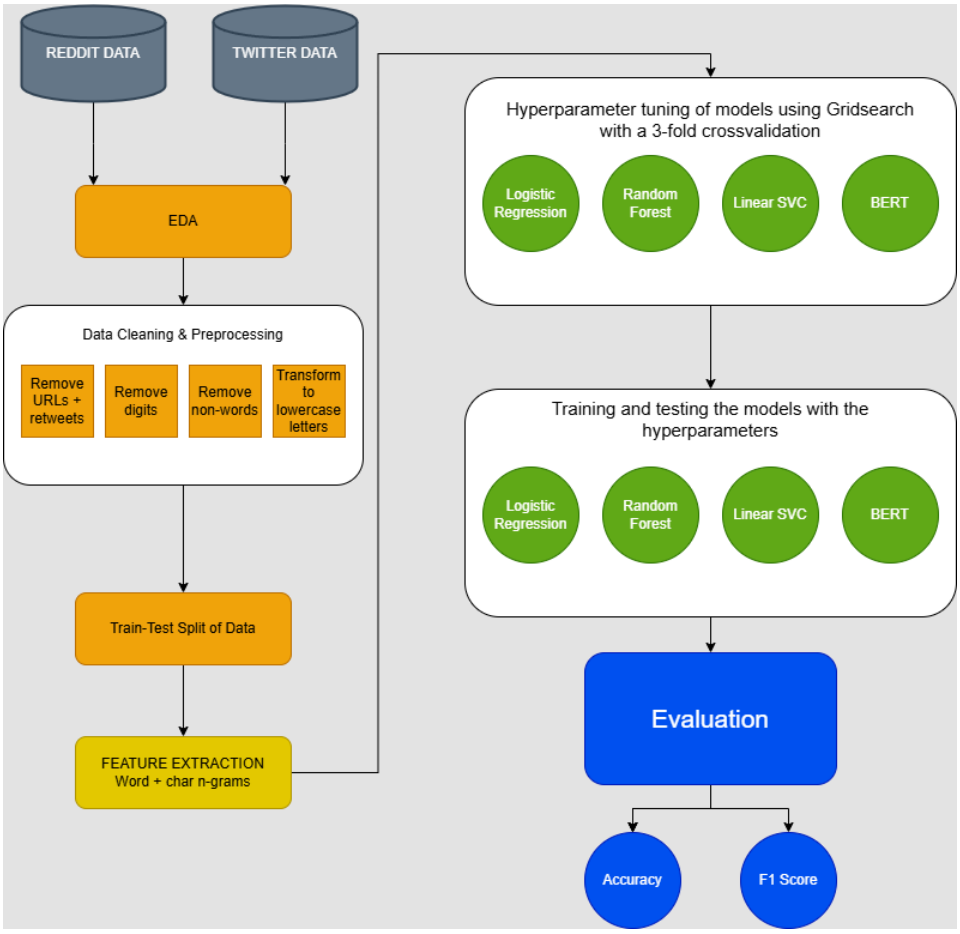

Figure 1: Flowchart Methodology

4.1 *Data*

The Reddit dataset has two years worth of Reddit history, from 2020 to 2022 and created by Emmery et al. (2024). Authors are labeled predominantly by the flairs they use, with gender being extracted from self-reports in posts, which are either male or female. Personality labels are based on self-reported MBTI types on their respective subreddits, such as r/enfp or r/intp. The gender data consists of 2401 unique authors and a total of 44,634 posts. The dataset is divided into several files: one dedicated to gender data and four separate files for the MBTI dimensions (Extraversion/Introversion, Sensing/iNtuitive, Thinking/Feeling, Judging/Perceiving).

Figure 2 provides a summary of the Reddit data, highlighting the post count and the number of unique users for each gender. The figure shows a clear balance between the gender classes.

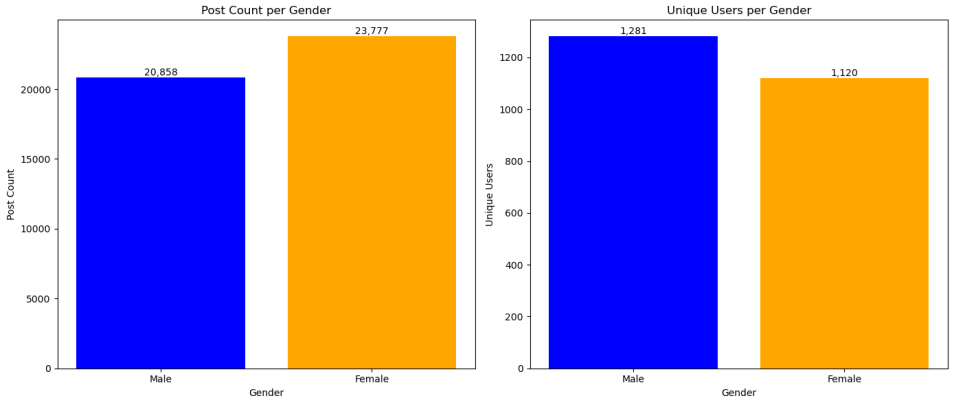

Figure 2: Reddit dataset information

The second dataset consists of tweets from users on Twitter, collected by Verhoeven et al. (2016). This dataset includes a total of 2,788,177 tweets associated with multiple users. Users are assigned with both gender, being either male or female, and MBTI labels, providing a comprehensive dataset for analysis.

The Twitter data contains significantly more tweets than the Reddit dataset. Figure 3 gives an overview of some key statistics of the Twitter dataset, including tweet counts and the number of unique users for each gender. The figure indicates a slight imbalance in the dataset.

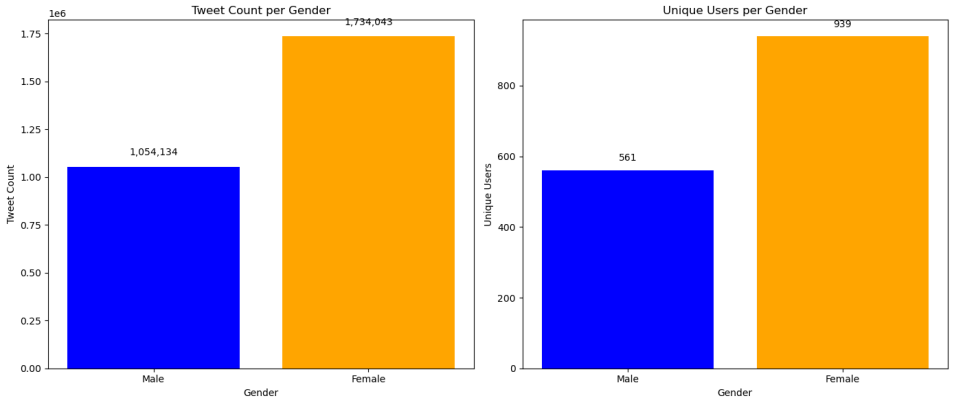

Figure 3: Twitter dataset information

It is important to check the distribution of the MBTI personality traits, to check how balanced they are. Firstly, the distribution of the MBTI types of Reddit data are displayed in figure 4. The distribution is displayed by checking how many posts are labeled to each corresponding MBTI type. This shows that the E/I and S/N dimensions are very imbalanced.

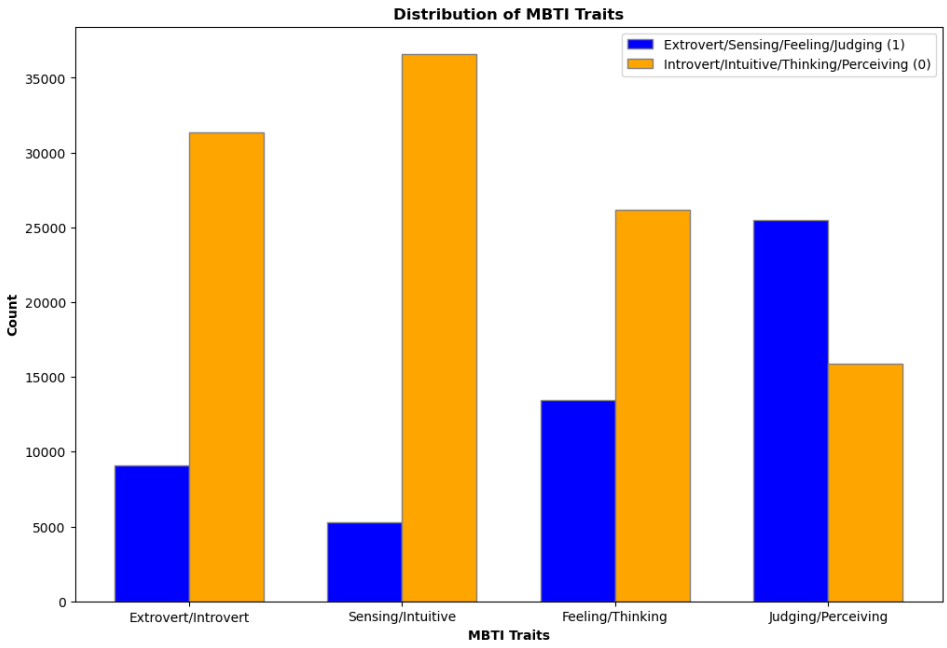

Figure 4: Reddit data MBTI class distribution

Figure 5 shows the MBTI type class distribution of the Twitter dataset. This one looks more balanced compared to the Reddit dataset. Here the distribution is also calculated by checking how many tweets are associated with a certain MBTI type. The figure shows that most dimensions are fairly balanced, with only the S/N dimension being imbalanced.

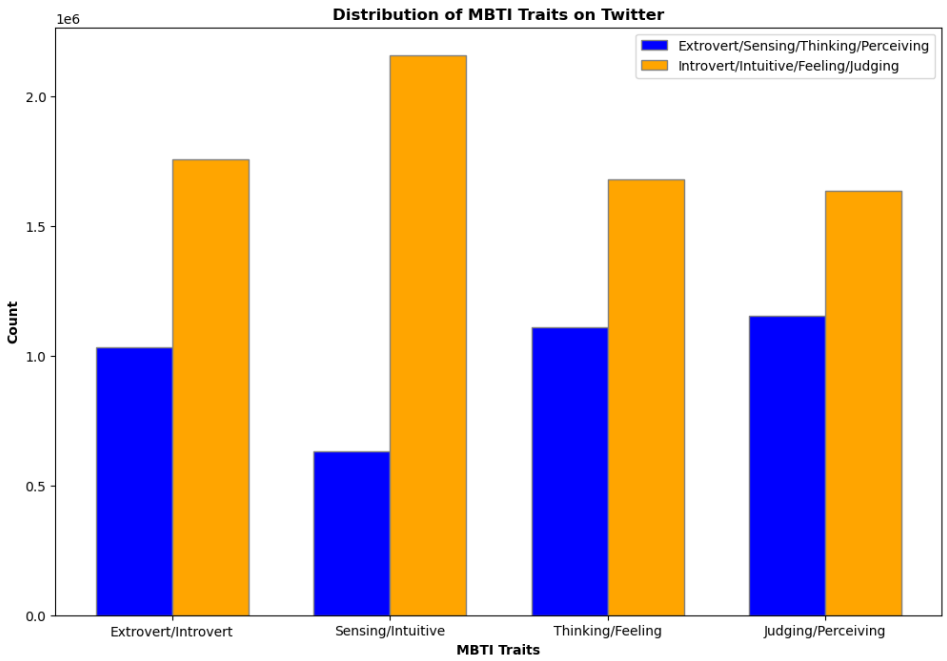

Figure 5: Twitter data MBTI class distribution

4.2 *Preprocessing*

After having these two datasets available and having checked the class distributions, the next part is to preprocess the data. Especially in the Twitter dataset, there are numerous tweets which contain interactions with other people, citations or 'retweets'. The goal is to remove these from the data. For the Reddit dataset, all non-words and digits have been removed and everything is transformed to lowercase letters. The removal of all non-words and digits is done in the same way for the Twitter data, hereafter the words are also converted to lowercase letters. Hereby the preprocessing of both datasets is done in the same way.

Term Frequency-Inverse Document Frequency, or TF-IDF, is an algorithm that is widely used in many NLP tasks (Vashisth & Meehan, 2020). This method is used to find the importance of a word in textual data or a dataset. The main use of the method is seeing how often a word or phrase occurs in the data. Hereby being suitable for classification (Liu et al., 2018). The TF-IDF method has been used in combination with the models in this research. A basic TF-IDF setup has been used for all models, with the same parameters. This has been done with the combination of the use of word and character n-grams. Character n-grams are commonly used text classification problems (Kruczek et al., 2020). Word n-grams are sequences of 'n' following words from a given text, while character n-grams

are sequences of 'n' following characters from a given text. Both word and character n-grams have proven to be one of the strongest predictor in gender classification tasks (Kunneman et al., 2017).

The Twitter dataset was a bigger dataset with more entries than the Reddit dataset as can be seen in Section 4.1. This was causing longer runtimes for the models. Because the time a model would need to process all this data and the amount of models being trained in this research, the Twitter data has been down-sampled for faster runtimes of the models. This has been done by using 10% of the original data, which still means a bigger dataset than the Reddit dataset. Since the amount of unique users is low, down-sampling has been done in a way to prevent losing any of these users. Hereby all the unique users are retained within the dataset and ensuring the diversity of the original dataset. Figure 6 shows the amount of unique users per gender, as well as the tweet count per gender for the Twitter data after being down-sampled and preprocessed. This shows that the unique users are retained, but the tweet count is lowered. The Reddit datasets still contains the same amount of users per gender and post count, because it was not down-sampled.

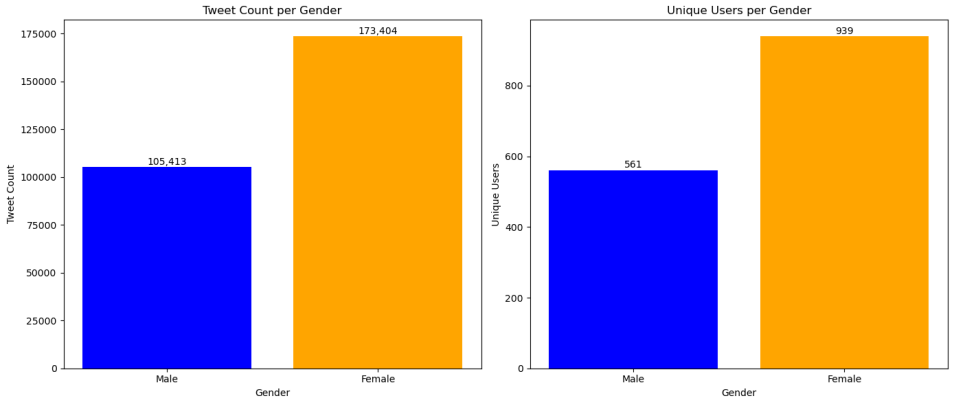

Figure 6: Twitter dataset information after down-sampling

After down-sampling the Twitter dateset, the MBTI distributions still remain similar to before down-sampling. Figure 7 shows the class distribution for the different MBTI types after down-sampling.

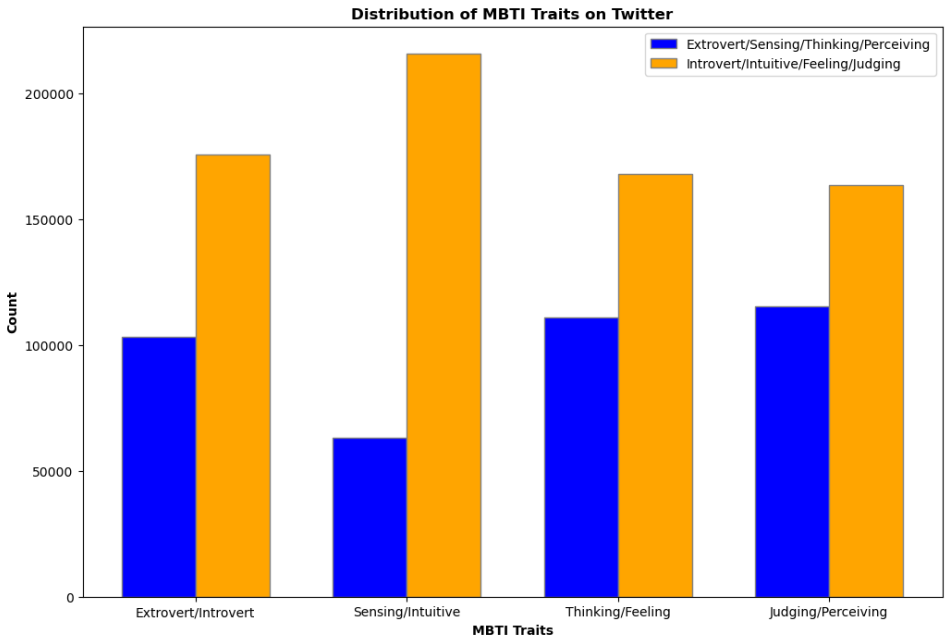

Figure 7: Twitter MBTI class distribution after down-sampling

4.3 *Models*

Multiple different machine learning models have been made to predict both gender and MBTI. These models will be outlined in this section. Models used in this research are Logistic Regression, Random Forest, Linear SVC and BERT. Logistic Regression is a fairly simple model, which is used in many binary classification tasks. Therefore it is one of the models used. Random Forest is an ensemble method which combines multiple different decision trees to improve the generalization of the model. These models have proven to be performing good over the years (Palomino-Garibay et al., 2015). The Linear SVC model is a Support Vector Machine (SVM). Support Vector Machines have proven to be generally effective of high dimensional data and are capable to accurately predict the samples. The Linear SVC uses a 'one-vs-all' strategy, meaning it trains as many models as there are decision classes (Lakhotia & Bresson, 2018). As seen in previous research by Lain and Zalzala (2023), Linear SVC was consistent on both single domain and cross-domain, making it a good model to use in this research. Finally, the BERT model. The BERT model is a relatively new model. It was introduced by Devlin et al. (2019). BERT stands for Biderectional Encoder Represenations from Tranformers. Unlike traditional unidirectional models, BERT uses a transformer to process text in a fully bidirectional manner, thus being able to capture richer context. BERT is a pre-trained model on

multiple text corpora. It can then be fine-tuned with minimal changes to be usable for a variety of NLP tasks. There are several variants of BERT models available, which are designed for different use cases. In this thesis, 'bert-base-uncased' has been used. This is a general-purpose model for English, and has also been used in previous works as can be found in Section 3.

The performance of the models will be compared to the heuristic baseline. This is a prediction strategy where the model would always predict the majority class. By comparing models to this heuristic baseline it would provide a simple yet effective benchmark to compare the models used in this thesis.

Every dimension in the MBTI indicator is treated separately, causing it to be four binary classification tasks. Previous literature has dealt with MBTI prediction in the same way and has proven to be effective. By doing so it can also give insights in which dimensions are more easily classified compared to the other dimensions. Gender is also a binary classification tasks with either male or female as label. It should be noted that for societal and ethical importance, gender is not binary, but this was not reflected in the data. The datasets had gender labeled as either male or female, which is why this research focused on either male or female.

For both the Reddit and Twitter dataset, the data has been split up by using 80% as training data and 20% as test data. This has been done on author-level. Meaning 80% of authors are in the training set and 20% of the authors are in the test set.

4.4 *Hyperparameter tuning*

All models have different hyperparameters available for tuning the models to get the optimal performance. Without tuned settings, models usually do not perform at optimal performance. In order to get the optimal hyperparameters for the different models, GridSearch has been used. GridSearch tries all the possible combination of parameters settings set by the user and gives the optimal settings. This has been done by using a cross-validation of 3, meaning every combination will be tested three times on different parts of the data. By using cross-validation within the GridSearch, the reliability of the model performance estimation is significantly improved. Cross-validation ensures that the parameters for a certain model are not only optimized for a single train-test split, but is optimized across multiple different subsets. The model's parameters have been optimized by running GridSearch in combination with the word and character n-grams.

Only the hyperparameters for Logistic Regression, Random Forest and Linear SVC have been optimized using GridSearch. The results of the

GridSearch for gender classification on Twitter and Reddit can be found in table 1. BERT has not been optimized, Keh, Cheng, et al. (2019) shows that tuning the learning rate, max sequence length and epochs only results in a minimal difference, where epochs had the biggest impact. Therefore the BERT model was ran using the parameters in table 1. Since the gender classes of the twitter dataset are slightly imbalanced, the models are built with the 'class_weight' function within the models. By setting this to 'balanced', it helps the model to handle imbalanced datasets.

Table 1: GridSearch Parameters for Gender Classification on Reddit and Twitter

Model	Parameters Values	Optimal Parameter Value
Logistic Regression	word ngram: [(1,2), (1,3)] char ngram: [(2,4), (2,5)] C: [0.1, 1, 10] penalty: [l2] svd: [100, 200]	word ngram: (1,3) char ngram: (2,4) C: 0.1 penalty: l2 svd: 200
Random Forest	word ngram: [(1,2), (1,3)] char ngram: [(2,4), (2,5)] n_estimators: [100, 200] max_depth: [10, 20, None]	word ngram: (1,2) char ngram: (2,5) n_estimators: 200 max_depth: None
Linear SVC	word ngram: [(1,2), (1,3)] char ngram: [(2,4), (2,5)] penalty: L2 C: [0.1, 1, 10]	word ngram: (1,2) char ngram: (2,4) penalty: L2 C: 0.1
BERT	epochs: [3] learning rate: [2e-5] max_seq_length: [128] batch size: [32]	epochs: 3 learning rate: 2e-5 max_seq_length: 128 batch size: 32

For the MBTI classification task, the tuning was done in the same way as for gender classification. Table 2 shows the results of the GridSearch. Here BERT has not been optimized with GridSearch, for the same reason as explained. Namely due to previous literature showing minimal to no positive effect and GridSearch consuming a lot of time. For the same reason as in the gender classification tasks, the models are built with the 'class_weight' function within the models. By setting this to 'balanced', it helps the model to handle imbalanced datasets.

Table 2: Gridsearch Parameters for MBTI Classification on Reddit and Twitter

Model	Parameter values	Optimal parameter value
Logistic Regression	word ngram: [(1,2), (1,3)] char ngram: [(2,4), (2,5)] C: [0.1, 1, 10] penalty: [l2] svd: [100, 200]	word ngram: (1,3) char ngram: (2,4) C: 0.1 penalty: l2 svd: 200
Random Forest	word ngram: [(1,2), (1,3)] char ngram: [(2,4), (2,5)] n_estimators: [100, 200] max_depth: [10, 20, None]	word ngram: (1,2) char ngram: (2,5) n_estimators: 200 max_depth: None
Linear SVC	word ngram: [(1,2), (1,3)] char ngram: [(2,4), (2,5)] penalty: L2 C: [0.1, 1, 10]	word ngram: (1,2) char ngram: (2,4) penalty: L2 C: 0.1
BERT	epochs: 3 learning rate: 2e-5 max_seq_length: 128 Batch size: 32	epochs: 3 learning rate: 2e-5 max_seq_length: 128 Batch size: 32

The same procedure is followed for the GridSearch on the cross-domain classification tasks. Table 3 shows the results of the GridSearch for cross-domain gender classification. BERT is again treated in the same way as for the single domain tasks.

Table 3: Gridsearch Parameters for cross-platform gender Classification

Model	Parameter values	Optimal parameter value
Logistic Regression	word ngram: [(1,2), (1,3)] char ngram: [(2,4), (2,5)] C: [0.1, 1, 10] penalty: [l2] svd: [100, 200]	word ngram: (1,3) char ngram: (2,4) C: 1 penalty: l2 svd: 200
Random Forest	word ngram: [(1,2), (1,3)] char ngram: [(2,4), (2,5)] n_estimators: [100, 200] max_depth: [10, 20, None]	word ngram: (1,2) char ngram : (2,5) n_estimators: 200 max_depth: None
Linear SVC	word ngram: [(1,2), (1,3)] char ngram: [(2,4), (2,5)] penalty: L2 C: [0.1, 1, 10]	word ngram: (1,2) char ngram: (2,4) penalty: L2 C: 0.1
BERT	epochs: 3 learning rate: 2e-5 max_seq_length: 128 Batch size: 32	epochs: 3 learning rate: 2e-5 max_seq_length: 128 Batch size: 32

Finally the GridSearch for the cross-domain classification for MBTI. These are visualized in table 4. As can be seen in all results of the tables, the changes in optimal parameters compared to single-domain differ slightly. BERT however, is treated in the same way as before, as can be seen in the table.

Table 4: Gridsearch Parameters for cross-platform MBTI Classification

Model	Parameter values	Optimal parameter value
Logistic Regression	word ngram: [(1,2), (1,3)] char ngram: [(2,4), (2,5)] C: [0.1, 1, 10] penalty: [l2] svd: [100, 200]	word ngram: (1,3) char ngram: (2,4) C: 1 penalty: l2 svd: 200
Random Forest	word ngram: [(1,2), (1,3)] char ngram: [(2,4), (2,5)] n_estimators: [100, 200] max_depth: [10, 20, None]	word ngram: (1,2) char ngram : (2,5) n_estimators: 200 max_depth: None
Linear SVC	word ngram: [(1,2), (1,3)] char ngram: [(2,4), (2,5)] penalty: L2 C: [0.1, 1, 10]	word ngram: (1,2) char ngram: (2,4) penalty: L2 C: 0.1
BERT	epochs: 3 learning rate: 2e-5 max_seq_length: 128 Batch size: 32	epochs: 3 learning rate: 2e-5 max_seq_length: 128 Batch size: 32

4.5 *Evaluation Metrics*

To evaluate the performance of the models, accuracy will be one of the metrics. Accuracy is a straightforward evaluation metric used in classification tasks, displaying the ratio of correctly predicted instances compared to the total amount of instances. Accuracy is a simple metric which is easy to understand and interpret. It provides a measure of how often a model's prediction matches the actual class labels. Accuracy is widely used in the field of author profiling. By using accuracy as performance metric, it aligns with the previous literature as seen in section 3, since many studies use this as metric for gender and MBTI classification tasks. Hereby making it an reliable metric.

To have more insight about the performance of the models, precision, recall and F1 score will also be evaluated. By also evaluating these metrics, the actual effectiveness of the models can be considered. These can complement accuracy as a metric by offering a more detailed view of how often a model predicts a certain class. This ensures a more robust model evaluation.

4.6 *Model packages and libraries*

All models and preprocessing steps are coded with Python. The different Python libraries and models used are the following: Scikit-learn (Pedregosa et al., 2018), Transformer (Vaswani et al., 2023), NumPy (Harris et al., 2020), Pandas (Mckinney, 2011), Matplotlib (Hunter, 2007), Tensorflow (Abadi et al., 2016), Logistic Regression (Maalouf, 2011), SVMs (Cortes & Vapnik, 1995), Random Forest (Breiman, 2001), BERT (Devlin et al., 2019).

5 RESULTS

Firstly single domain results will be reported for both gender and MBTI classification, by reporting the accuracy and macro-F1 scores. Afterwards the cross-domain performance of the different models will be reported in the same manner.

5.1 *Single domain results*

5.1.1 *Single domain gender classification*

Taking a look at the Reddit gender classification, it is clear that the accuracy is fairly high, with the lowest accuracy being 0.78 and the highest being 0.91. Results of the gender classification on both Reddit and Twitter can be found in table 5. The table shows the accuracy of the models and the macro-F1 scores, with the highest accuracy and F1 score of either dataset highlighted in bold. The best performing model was Logistic Regression for predicting gender on the Reddit dataset, with an accuracy of 0.91, which also had a macro-F1 score of 0.91, which indicates a balanced and effective classification. However, on the Twitter dataset it can be seen as the worst performing model, with an accuracy of only 0.56 and the same F1 score. For the Twitter data, the Linear SVC is the best performing model with an accuracy of 0.76 and a F1 score of 0.75, while the same model is one of the worst performing models on the Reddit dataset. The results show a drop in performance of the models on Twitter compared to Reddit.

Table 5: Model performance for gender classification on Reddit & Twitter

	Reddit		Twitter	
Model	Accuracy	F1 Score	Accuracy	F1 Score
Logistic Regression	**0.91**	**0.91**	0.56	0.56
Random Forest	0.84	0.84	0.62	0.47
Linear SVC	0.78	0.78	**0.76**	**0.75**
BERT	0.78	0.76	0.64	0.64
Majority Baseline	0.53	0.35	0.62	0.38

While accuracy and macro-F1 scores are visible in table 5, more metric are available in Appendix A(page 34). The classification reports are reported there, including the precision, recall and F1 scores for the models on both Reddit and Twitter data.

5.1.2 *Single domain MBTI classification*

The models have predicted MBTI in a way that it consists of four different binary classification tasks. Every dimensions in the MBTI personality type has been predicted separately. Hereby the results are based on a single dimension, so for example the Extraversion/Introversion (E/I) dimension. Here the accuracy is calculated on the basis how often it correctly predicted one of the two classes. The result of MBTI prediction on the Reddit data set in found in table 6. The table shows the accuracy of the models and the macro-F1 scores of the models on the different dimensions. From these results it can be seen that the Linear SVC model is scoring high on all different dimensions, with a minimum accuracy of 0.86 for the J/P dimension and achieving the highest accuracy on the S/N dimension with an accuracy of 0.95. This is paired with high F1 scores, ranging from 0.84 to 0.87, indicating a reliable model. The results indicate BERT as the worst performing model on this task. While Random Forest has decent accuracy on most dimensions, the lower F1 scores indicate it is more biased towards the majority class. As the table indicates, all models achieve much higher F1 scores compared to the baseline model.

Table 6: Model performance for MBTI classification on Reddit

	Accuracy				F1 Score			
Model	E/I	S/N	T/F	J/P	E/I	S/N	T/F	J/P
Logistic Regression	0.86	0.93	0.83	0.80	0.78	0.83	0.81	0.79
Random Forest	0.81	0.90	0.77	0.75	0.56	0.68	0.68	0.70
Linear SVC	**0.90**	**0.95**	**0.88**	**0.86**	**0.84**	**0.87**	**0.87**	**0.86**
BERT	0.79	0.71	0.68	0.74	0.52	0.71	0.72	0.78
Majority Baseline	0.78	0.87	0.66	0.62	0.44	0.47	0.40	0.38

The results for the Twitter data are different than for the Reddit data. As can be seen in table 7. The table provides the accuracy and macro-F1 scores of all models on the different MBTI dimensions. All models score almost equally on all different dimensions based on the accuracy. The top 2 performing models based on accuracy are Random Forest and Linear SVC. Only the Logistic Regression model is scoring slightly lower, while BERT is the worst performing model. However the F1 scores are lower for Random Forest compared to Linear SVC. Based on the F1 scores, Logistic Regression is showing the best performance. With accuracy of the S/N dimension being fairly high across all models, the low F1 score indicates issues with precision and recall of the models on the two different dimensions. Linear SVC and Logistic Regression show the best performance with balanced accuracy and F1 scores. Nearly all models outperform the majority baseline model on accuracy on all dimensions, as do they greatly score better on the F1 score.

Table 7: Model performance for MBTI classification on Twitter

	Accuracy				F1 Score			
Model	E/I	S/N	T/F	J/P	E/I	S/N	T/F	J/P
Logistic Regression	0.64	**0.80**	**0.64**	0.62	**0.58**	**0.52**	**0.58**	**0.56**
Random Forest	**0.65**	**0.80**	**0.64**	**0.63**	0.54	0.51	0.54	0.45
Linear SVC	**0.65**	**0.80**	**0.64**	**0.63**	**0.58**	0.51	0.57	**0.56**
BERT	0.59	0.73	0.60	0.61	0.54	0.45	0.49	0.54
Majority Baseline	0.63	0.77	0.60	0.59	0.39	0.44	0.38	0.37

Since both accuracy and macro-F1 scores are displayed in both table 6 and table 7, the classification reports are available in Appendix B(page 35). Here the precision, recall and F1 scores are visualized in an overview for the MBTI classification on Reddit and Twitter data.

5.2 *Cross-domain results*

Following on the previous results which are on a single domain, the cross-domain results will be highlighted in this section.

5.2.1 *Cross-domain Gender Classification*

Table 8 shows the results for gender classification with models trained and tested on the two datasets, with the highest scores highlighted. It displays the accuracy and macro-F1 scores for the models. Firstly looking at results from Twitter as training data and Reddit as testing data, Logistic Regression achieved the highest accuracy of 0.65. Logistic Regression also achieved the highest F1 score of 0.64, however only slightly higher than Linear SVC. With Reddit as training data and Twitter as testing data, Linear SVC achieved the highest accuracy and F1 score, scoring 0.59 and 0.57 respectively. Random Forest appears as the worst performing model, with only a F1 score of 0.36 with Reddit as testing data, indicating issues with precision and recall on the two different classes. BERT shows moderate performance across both domains. Hereby indicating that the Linear SVC and Logistic Regression models demonstrate the most robust performance across both domains and outperforming the baseline model.

Table 8: Cross-platform model performance for gender classification

	Train:Twitter Test:Reddit		Train:Reddit Test:Twitter	
Model	Accuracy	F1 Score	Accuracy	F1 Score
Logistic Regression	**0.65**	**0.64**	0.53	0.53
Random Forest	0.48	0.36	0.52	0.51
Linear SVC	0.64	0.63	**0.59**	**0.57**
BERT	0.56	0.56	0.54	0.54
Majority Baseline	0.62	0.38	0.53	0.35

More detailed classification reports on the cross-platform model performance for gender classification can be found in Appendix C(page 37).

5.2.2 *Cross-domain MBTI Classification*

Now reviewing the cross-platform performance of the models for MBTI classification. Firstly, models trained on Twitter and tested on Reddit data. The results can be seen in table 9, with the highest scores highlighted in bold. It shows the accuracy and the macro-F1 scores of the models on the four different MBTI dimensions. BERT displays the highest accuracy across the E/I, T/F and J/P dimensions, while Random Forest excels in the S/N

dimension. However, the F1 scores for Random Forest are significantly lower across all dimensions, except for the J/P dimension where it has the highest F1 score, indicating issues with classifying the different classes on the other dimensions. Logistic Regression and Linear SVC perform similar according to the results. Overall, BERT is showing best performance across different dimensions, except on the S/N dimension.

Table 9: Cross-platform model performance for MBTI classification using Twitter as training data and Reddit as testing data

	Accuracy				F1 Score			
Model	E/I	S/N	T/F	J/P	E/I	S/N	T/F	J/P
Logistic Regression	0.67	0.85	0.63	0.63	**0.51**	**0.51**	0.45	0.46
Random Forest	0.77	**0.87**	0.66	0.55	0.44	0.47	0.42	**0.49**
Linear SVC	0.68	0.86	0.54	0.63	**0.51**	0.50	0.43	0.46
BERT	**0.79**	0.72	**0.68**	**0.67**	0.50	0.48	**0.48**	0.44
Majority Baseline	0.63	0.77	0.60	0.59	0.39	0.44	0.38	0.37

Lastly, the results of models trained on Reddit data and using Twitter as test data on MBTI. The performance of the models can be found in table 10. As can be seen, the baseline model achieved the highest accuracy across all dimensions. However, the F1 scores are much lower. The class distribution of the Reddit dataset for MBTI had shown the class-imbalances in section 4.1. If the baseline model is left out, BERT stands out with the highest accuracy on the E/I, T/F and J/P dimensions, yet the F1 scores remain relatively low. Logistic Regression and Linear SVC have better results on the F1 score compared to the other models. While Random Forest has the highest accuracy on the S/N dimension, it has significantly lower F1 scores compared to other models, indicating problems with identifying the two different classes.

Table 10: Cross-platform model performance for MBTI classification using Reddit as training data and Twitter as testing data

	Accuracy				F1 Score			
Model	E/I	S/N	T/F	J/P	E/I	S/N	T/F	J/P
Logistic Regression	0.52	0.70	0.53	0.50	**0.49**	**0.51**	**0.51**	**0.52**
Random Forest	0.57	0.80	0.45	0.52	0.40	0.45	0.25	0.34
Linear SVC	0.52	0.71	0.52	0.48	**0.49**	**0.51**	**0.51**	0.51
BERT	0.63	0.72	0.58	0.58	0.39	0.48	0.42	0.39
Majority Baseline	**0.78**	**0.87**	**0.66**	**0.62**	0.44	0.47	0.40	0.38

The metrics as precision, recall and F1 scores, which come from the classification reports of the models can be found in Appendix D(page 38).

6 DISCUSSION

The goal of this research was to evaluate different machine learning models on performing classification tasks, namely gender and MBTI personality type prediction of an author on social media platforms. More specifically of authors on Reddit and Twitter. Firstly on single domain basis, where models were trained on Twitter data and tested on the same data. Afterwards to see how these models would perform cross-domain, so trained on Reddit data and tested on Twitter data. This was also done the other way around, meaning models were trained on Twitter and tested on Reddit to see the difference in performance of models trained and tested on different datasets. The goal was to see the performance of models and how performance would change when they would be trained and tested on both Reddit and Twitter data.

6.1 *Results Discussion*

6.1.1 *Single domain discussion*

When looking at the gender classification tasks on Reddit and Twitter data, Linear SVC was the highest scoring model on Twitter, with an accuracy of 0.76. While it is one of the lowest scoring models on Reddit. The highest accuracy was 0.82 in the 2018 PAN task for gender classification on Twitter data (Rangel et al., 2018). With the highest scoring model on Reddit being 0.91, it is an improvement, but on another social media paltform. However the results on Twitter perform worse and contradicts with prior research of Rangel et al. (2018), where Logistic Regression and SVM were the highest scoring models. Logistic regression achieved the lowest score on our Twitter dataset. The highest scoring model in the 2018 PAN task made use of more extensive feature engineering, while the feature engineering and preprocessing in this research was kept limited. However, their model achieved an accuracy of 0.82 with Linear SVC and the Linear SVC in this thesis achieved an accuracy of 0.76 with a respectable F1 score of 0.75, with less preprocessing and feature engineering steps. Meaning Logistic Regression is more reliant on better preprocessing and feature engineering for better performance on Twitter data. Overall the results on Reddit for gender classification perform well above the majority baseline model.

Considering the MBTI predictions, models on Reddit data scored fairly high. With Nisha et al. (2022) having the scored highest on the S/N

dimensions, is aligned with our results where S/N is also the highest scoring dimension across all model, for both Reddit and Twitter data. Although the S/N dimension being one of the dimensions with the highest accuracy, Plank and Hovy (2015) have seen the majority baseline model also achieving good accuracy. In our Twitter and Reddit dataset, the S/N dimension had the biggest class imbalance, which is a cause for higher accuracy and therefore lower F1 scores. Overall, Linear SVC is the best scoring model on Reddit and one of the better models on Twitter data. The models score lower on Twitter data compared to Reddit data. This can be because of the down-sampling, which was done on 'tweet' level. This can affect the quality of the data. Another possibility could be the difference in how users express themselves on Twitter compared to Reddit, which means better preprocessing is needed. All models seem to perform better than the majority baseline model overall.

6.1.2 *Cross-domain discussion*

The cross-domain results show that performance of the models are decreasing. This is to be expected, since models usually do not generalize well on unseen data and every social media platform data is different. By firstly looking at the results on gender classification, we see that the accuracy is lower in general across all models. The highest accuracy on Reddit data only was 0.90 with Logistic Regression, while the accuracy on Reddit as test data dropped to 0.65 with Logistic Regression. On Twitter data only was 0.76 by using the Linear SVC model, with accuracy dropping to 0.59 on Twitter as test data, also with a Linear SVC model. Interesting to see, is that the model that had the highest accuracy on a single domain, is also the model with the highest accuracy by using different training data than test data. Linear SVC showing consistent performance on both single and cross-domain is in line with research done by Lain and Zalzala (2023), where Linear SVC also had consistent results. Reasoning for lower performance is for the fact that the models have issues generalizing on unseen data. Results indicate models can predict better on Reddit data compared to Twitter data, which shows Twitter data needs more preprocessing steps to achieve better performance for the models. However, the best performing models still outperform the majority baseline model, indicating good progress, but there is still room for improvement.

Now taking a look at the results for cross-platform MBTI classification. On the Reddit data only, Linear SVC was clearly the best performing model across all different dimensions. By seeing the results of the models with Twitter as training data and Reddit as testing data, we see that Linear SVC is no longer the top performing model. BERT is achieving the highest accuracy on the E/I, T/F, and J/P dimensions, while having the lowest

accuracy on the S/N dimension. This is in line with a previous study by Keh, Cheng, et al. (2019) and Verhoeven et al. (2016), where BERT is performing best on the E/I and T/F dimensions. Random Forest achieved the highest accuracy on the S/N dimension for both cross-domain tests, but shows a significant drop in F1 score. This indicates problems with identifying the two different classes, which is caused by the class imbalance in the S/N dimension. This is in line with previous research done by Plank and Hovy (2015). BERT was not the best performing model on single domain tasks, but is performing well on cross-domain tasks. It achieves the highest accuracy on both Twitter as training and Reddit as test data, as well as Reddit as training and Twitter as test data. However, with Reddit as training data and Twitter as testing data, BERT shows a drop in F1 scores, whereas Logistic Regression and Linear SVC achieve better F1 scores. Indicating BERT has issues with predicting on Twitter data. While one reason could be again due to not enough preprocessing, Alzahrani and Jololian (2021) conducted research on the effect of preprocessing steps on data before using a BERT model for prediction tasks, resulting in BERT performing best with no preprocessing of the data.

6.2 *Limitations*

There are several limitations in this study that should be acknowledged. Firstly, the social media datasets. While the Reddit and Twitter datasets are good datasets, the Twitter datasets was down-sampled due to time constraints. This decision was made to reduce the runtime for models, but has an effect on the performance, which can be a reason for lower performance of models on the Twitter dataset. Also the class-imbalances seem to have had some impact, especially on the MBTI predictions. While some measures had been taken, another form of resampling should be evaluated. Although, as per the gender prediction which is a balanced dataset, model performance was still not optimal and can be improved. Additionally, the preprocessing steps and feature engineering steps have been kept limited in this research. This study was focused on the model's capability of predicting on unseen data of another social media platform. By focusing more on feature engineering and preprocessing of the data, the performance of models can possibly be improved. Lastly, only four different machine learning models have been evaluated. With Logistic Regression, Linear SVC and BERT showing decent performance across both single and cross-domain, other models can also prove to be effective.

6.3 *Relevance*

As mentioned before, Reddit is a lesser used social media platform used in author profiling tasks, compared to a social media platform like Twitter. While also being more relevant lately in the literature, the literature on cross-domain usability of models is still limited. This study explored the performance of models on both single domain and cross-domain on social media platforms for both gender and MBTI classification tasks. Hereby this study contributes to the use of models for cross-platform usability, when a certain social media data can be scarce, the use of other social media data could be used to train models and predict on the other platform. Personality traits prediction can be used for multiple different reason. Firstly, content recommendation which could enhance user experience on a platform. Secondly, targeted marketing campaigns for companies, this could however raise ethical concerns about user privacy. Finally, as studied by Preoţiuc-Pietro et al. (2015), personality prediction on social media can help diagnose psychological disorder of users.

6.4 *Future work*

Despite using four different models, including a more advanced model like BERT, performance is based on different factors. There are a few possibilities for follow-up research on this topic. Firstly, it would be valuable to research how different feature engineering and preprocessing methods could change the performance of these models. As researched by Alzahrani and Jololian (2021), BERT performs best with no preprocessing. Future work could research the performance of the different models by changing preprocessing and feature engineering methods for cross-platform generalizability.

Furthermore, BERT showed better performance for cross-domain usability. BERT is a transfer learning model, but there are multiple different models comparable to BERT available, like XLNet. Research could explore how other models similar to BERT perform on author profiling tasks.

Lastly, this study has used two different social media datasets, namely Reddit and Twitter. By using multiple different social media datasets to train a model and evaluate the performance on another social media dataset excluded from the training, could provide more insights into the cross-domain usability of models. Models could possibly perform better when trained on multiple different datasets to finally predict on an unseen dataset.

7 CONCLUSION

The main goal of this study was to see the performance of models on single domain and cross-domain on gender and MBTI classification tasks for Reddit and Twitter. This section will answer the research questions of this study.

SUB-RQ1 *Which of the models(LR, RF, Linear SVC and BERT) are most accurate on only Reddit or Twitter data?*

The best performing model for gender classification on Reddit data is Logistic Regression with an accuracy of 0.91 and F1 score of 0.91. Linear SVC is the best performing model on Twitter data with an accuracy of 0.76 and a F1 score of 0.75. Looking at the MBTI results, Linear SVC outperforms all models with accuracy ranging from 0.95 to 0.86 and F1 ranges from 0.87 to 0.84 on Reddit data. For Twitter data, Linear SVC and Random Forest achieve the highest accuracy, followed by Logistic Regression. Although the F1 scores for Random Forest are significantly lower, causing both Linear SVC and Logistic Regression to be the best performing models.

SUB-RQ2 *Does the cross-domain performance of models change if they are trained on Reddit or on Twitter data?*

Looking at the cross-domain performance of the models, the accuracy of models trained on Twitter and tested on Reddit data perform better then the other way around. This is the case for both gender classification and MBTI prediction. The preprocessing of the Twitter data can be a cause for lower predictive performance of the models on the data.

MAIN-RQ *How do machine learning models(LR, RF, Linear SVC and BERT) trained on Twitter data perform on Reddit data for author profiling tasks, with the focus on gender and MBTI personality traits?*

Based on the findings in this study, the performance of the machine learning models dropped cross-domain compared to single domain. For the gender classification task, Logistic Regression was the best performing model while trained on Twitter and tested on Reddit data, with Linear SVC performing only slightly worse. The results indicate that BERT achieved the highest accuracy across most dimensions for MBTI classification while being trained on Twitter and tested on Reddit. However, the F1 scores of Logistic Regression and Linear SVC also compare to the F1 scores of the BERT model. To conclude, the machine learning models show good potential in cross-domain usability, but there is still room for improvement considering the recommendations for future work and the limitations in this study.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., … Zheng, X. (2016). Tensorflow: A system for large-scale machine learning.

Alzahrani, E., & Jololian, L. (2021). How different text-preprocessing techniques using the bert model affect the gender profiling of authors (D. C. Wyld & D. Nagamalai, Eds.). https://doi.org/10.48550/arXiv.2109.13890

Ameer, I., Sidorov, G., & Nawab, R. M. A. (2019). Author profiling for age and gender using combinations of features of various types. *Journal of Intelligent & Fuzzy Systems*, *36*(5), 4833–4843.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Briggs Myers, I., & Myers, P. B. (2010). *Gifts differing: Understanding personality type*. Nicholas Brealey Publishing.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dias, R., & Paraboni, I. (2020). Cross-domain author gender classification in brazilian portuguese. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1227–1234.

dos Santos, V. G., & Paraboni, I. (2022). Myers-briggs personality classification from social media text using pre-trained language models. *J. Univers. Comput. Sci.*, *28*, 378–395. https://api.semanticscholar.org/CorpusID:248438282

Emmery, C., Chrupała, G., & Daelemans, W. (2017). Simple queries as distant labels for predicting gender on twitter. *Proceedings of the 3rd Workshop on Noisy User-generated Text*, 50–55.

Emmery, C., Miotto, M., Kramp, S., & Kleinberg, B. (2024). SOBR: A corpus for stylometry, obfuscation, and bias on reddit. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources, and Evaluation*.

Gjurković, M., Karan, M., Vukojević, I., Bošnjak, M., & Snajder, J. (2021, June). PANDORA talks: Personality and demographics on Reddit.

In L.-W. Ku & C.-T. Li (Eds.), *Proceedings of the ninth international workshop on natural language processing for social media* (pp. 138–152). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.socialnlp-1.12

Gjurković, M., & Šnajder, J. (2018, June). Reddit: A gold mine for personality prediction. In M. Nissim, V. Patti, B. Plank, & C. Wagner (Eds.), *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media* (pp. 87–97). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-1112

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with numpy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Keh, S. S., Cheng, I., et al. (2019). Myers-briggs personality classification and personality-specific language generation using pre-trained language models. *arXiv preprint arXiv:1907.06333*.

Kruczek, J., Kruczek, P., & Kuta, M. (2020). Are n-gram categories helpful in text classification? *Computational Science – ICCS 2020*, *12138*, 524–537. https://doi.org/10.1007/978-3-030-50417-5_39

Kunneman, F., Nguyen, D., Pardo, A., Brussee, M. W. P., & van der Plas, L. J. M. (2017). N-gram: New groningen author-profiling model. *arXiv preprint arXiv:1707.03764*. https://ar5iv.labs.arxiv.org/html/1707.03764

Lain, A. G., & Zalzala, A. M. S. (2023). A transfer learning approach to cross-domain author profiling. *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*.

Lakhotia, S., & Bresson, X. (2018). An experimental comparison of text classification techniques. *2018 International Conference on Cyberworlds (CW)*, 58–65.

Liu, C.-z., Sheng, Y.-x., Wei, Z.-q., & Yang, Y.-Q. (2018). Research of text classification based on improved tf-idf algorithm. *2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)*, 218–222.

Maalouf, M. (2011). Logistic regression in data analysis: An overview. *International Journal of Data Analysis Techniques and Strategies*, *3*, 281–299. https://doi.org/10.1504/IJDATS.2011.041335

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, *60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

Mckinney, W. (2011). Pandas: A foundational python library for data analysis and statistics. *Python High Performance Science Computer*.

Mirkin, S., Nowson, S., Brun, C., & Perez, J. (2015, September). Motivating personality-aware machine translation. In L. Màrquez, C. Callison-Burch, & J. Su (Eds.), *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1102–1108). Association for Computational Linguistics. https://doi.org/10.18653/v1/D15-1130

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2016). Survey: Computational sociolinguistics: A Survey. *Computational Linguistics*, *42*(3), 537–593. https://doi.org/10.1162/COLI_a_00258

Nisha, K. A., Kulsum, U., Rahman, S., Hossain, M. F., Chakraborty, P., & Choudhury, T. (2022). A comparative analysis of machine learning approaches in personality prediction using mbti. *Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2021*, 13–23.

Palomino-Garibay, A., Camacho-Gonzalez, A. T., Fierro-Villaneda, R. A., Hernandez-Farias, I., Buscaldi, D., Meza-Ruiz, I. V., et al. (2015). A random forest approach for authorship profiling. *Proceedings of CLEF*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2018). Scikit-learn: Machine learning in python.

Plank, B., & Hovy, D. (2015, September). Personality traits on Twitter—or—How to get 1,500 personality tests in a week. In A. Balahur, E. van der Goot, P. Vossen, & A. Montoyo (Eds.), *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 92–98). Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-2913

Preoţiuc-Pietro, D., Eichstaedt, J., Park, G., Sap, M., Smith, L., Tobolsky, V., Schwartz, H. A., & Ungar, L. (2015). The role of personality, age, and gender in tweeting about mental illness. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 21–30. https://doi.org/10.3115/v1/W15-1203

Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., & Stein, B. (2018). Overview of the 6th author profiling task at pan 2018: Multimodal

gender identification in twitter [Available online at CEUR Workshop Proceedings]. *Proceedings of the PAN 2018 Conference*.

Reddy, T. R., Vardhan, B. V., & Reddy, P. V. (2017). N-gram approach for gender prediction. *2017 IEEE 7th International Advance Computing Conference (IACC)*, 860–865. https://doi.org/10.1109/IACC.2017.0176

Vashisth, P., & Meehan, K. (2020). Gender classification using twitter text data. *2020 31st Irish Signals and Systems Conference (ISSC)*, 1–6.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need.

Verhoeven, B., Daelemans, W., & Plank, B. (2016, May). TwiSty: A multilingual Twitter stylometry corpus for gender and personality profiling. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 1632–1637). European Language Resources Association (ELRA). https://aclanthology.org/L16-1258

APPENDIX A. CLASSIFICATION REPORTS FOR GENDER ON SINGLE DOMAIN.

Table 11: Precision, Recall and F1 Score for Gender on Reddit data

Model	Class	Precision	Recall	F1 Score
Logistic Regression	M	0.91	0.92	0.92
	F	0.91	0.90	0.90
Random Forest	M	0.83	0.90	0.86
	F	0.87	0.79	0.83
Linear SVC	M	0.79	0.80	0.80
	F	0.77	0.75	0.76
BERT	M	0.78	0.78	0.78
	F	0.87	0.63	0.74

Table 12: Precision, Recall and F1 Score for Gender on Twitter data

Model	Class	Precision	Recall	F1 Score
Logistic Regression	M	0.45	0.61	0.52
	F	0.69	0.53	0.60
Random Forest	M	0.57	0.10	0.17
	F	0.63	0.95	0.76
Linear SVC	M	0.69	0.66	0.68
	F	0.80	0.82	0.81
BERT	M	0.54	0.68	0.60
	F	0.77	0.62	0.68

APPENDIX B. CLASSIFICATION REPORTS FOR MBTI ON SINGLE DOMAIN.

Table 13: Precision, Recall and F1 Score for MBTI Dimensions on Reddit

Dimension	Model	Class	Precision	Recall	F1 Score
E/I	Logistic Regression	E	0.89	0.94	0.91
		I	0.73	0.59	0.65
	Random Forest	E	0.80	1.00	0.89
		I	0.95	0.13	0.23
	Linear SVC	E	0.91	0.97	0.94
		I	0.88	0.64	0.74
	BERT	E	0.75	0.95	0.84
		I	0.92	0.12	0.20
S/N	Logistic Regression	S	0.94	0.98	0.96
		N	0.85	0.60	0.70
	Random Forest	S	0.90	1.00	0.95
		N	0.99	0.25	0.41
	Linear SVC	S	0.95	1.00	0.97
		N	0.96	0.63	0.76
	BERT	S	0.72	0.80	0.76
		N	0.80	0.56	0.65
T/F	Logistic Regression	T	0.86	0.89	0.87
		F	0.77	0.71	0.74
	Random Forest	T	0.75	0.98	0.85
		F	0.89	0.36	0.51
	Linear SVC	T	0.89	0.94	0.91
		F	0.86	0.77	0.82
	BERT	T	0.70	0.78	0.74
		F	0.77	0.68	0.70
J/P	Logistic Regression	J	0.77	0.71	0.74
		P	0.82	0.86	0.84
	Random Forest	J	0.88	0.42	0.57
		P	0.72	0.96	0.82
	Linear SVC	J	0.86	0.78	0.82
		P	0.86	0.92	0.89
	BERT	J	0.76	0.70	0.73
		P	0.81	0.84	0.83

Table 14: Precision, Recall and F1 Score for MBTI Dimensions on Twitter data

Dimension	Model	Class	Precision	Recall	F1 Score
E/I	Logistic Regression	E	0.58	0.33	0.42
		I	0.66	0.85	0.74
	Random Forest	E	0.63	0.23	0.34
		I	0.65	0.91	0.74
	Linear SVC	E	0.59	0.32	0.41
		I	0.66	0.86	0.75
	BERT	E	0.53	0.28	0.37
		I	0.61	0.80	0.70
S/N	Logistic Regression	N	0.80	0.99	0.89
		S	0.64	0.08	0.15
	Random Forest	N	0.80	0.99	0.89
		S	0.72	0.07	0.13
	Linear SVC	N	0.80	0.99	0.89
		S	0.72	0.07	0.13
	BERT	N	0.70	0.90	0.80
		S	0.60	0.07	0.10
T/F	Logistic Regression	F	0.66	0.84	0.74
		T	0.57	0.32	0.41
	Random Forest	F	0.64	0.92	0.76
		T	0.63	0.21	0.31
	Linear SVC	F	0.66	0.85	0.74
		T	0.58	0.31	0.40
	BERT	F	0.62	0.82	0.71
		T	0.55	0.23	0.28
J/P	Logistic Regression	J	0.64	0.85	0.73
		P	0.56	0.29	0.39
	Random Forest	J	0.63	0.92	0.75
		P	0.62	0.20	0.30
	Linear SVC	J	0.64	0.85	0.73
		P	0.57	0.29	0.38
	BERT	J	0.61	0.82	0.70
		P	0.54	0.28	0.38

APPENDIX C. CLASSIFICATION REPORTS FOR GENDER ON CROSS-DOMAIN.

Table 15: Precision, Recall and F1 Score for Gender on models trained on Twitter data and Reddit as test data

Model	Class	Precision	Recall	F1 Score
Logistic Regression	M	0.80	0.45	0.58
	F	0.58	0.87	0.70
Random Forest	M	0.88	0.04	0.07
	F	0.48	0.99	0.64
Linear SVC	M	0.79	0.44	0.56
	F	0.58	0.87	0.69
BERT	M	0.45	0.61	0.52
	F	0.69	0.53	0.60

Table 16: Precision, Recall and F1 Score for Gender on models trained on Reddit data and Twitter as test data

Model	Class	Precision	Recall	F1 Score
Logistic Regression	M	0.43	0.67	0.53
	F	0.68	0.43	0.53
Random Forest	M	0.50	0.70	0.59
	F	0.59	0.40	0.43
Linear SVC	M	0.49	0.78	0.60
	F	0.74	0.42	0.53
BERT	M	0.44	0.59	0.51
	F	0.66	0.50	0.57

APPENDIX D. CLASSIFICATION REPORTS FOR MBTI ON CROSS-DOMAIN.

Table 17: Precision, Recall and F1 Score for MBTI Dimensions on models trained on Twitter data and Reddit as test data

Dimension	Model	Class	Precision	Recall	F1 Score
E/I	Logistic Regression	E	0.23	0.21	0.22
		I	0.78	0.80	0.79
	Random Forest	E	0.00	0.00	0.00
		I	0.78	1.00	0.87
	Linear SVC	E	0.24	0.19	0.21
		I	0.78	0.82	0.80
	BERT	E	0.75	0.95	0.85
		I	0.20	0.08	0.15
S/N	Logistic Regression	N	0.88	0.96	0.92
		S	0.20	0.06	0.09
	Random Forest	N	0.87	1.00	0.93
		S	0.00	0.00	0.00
	Linear SVC	N	0.88	0.97	0.92
		S	0.18	0.05	0.08
	BERT	N	0.78	0.94	0.85
		S	0.24	0.06	0.10
T/F	Logistic Regression	F	0.31	0.07	0.12
		T	0.66	0.92	0.77
	Random Forest	F	0.27	0.03	0.06
		T	0.66	0.95	0.78
	Linear SVC	F	0.30	0.05	0.08
		T	0.66	0.94	0.78
	BERT	F	0.24	0.06	0.10
		T	0.78	0.94	0.85
J/P	Logistic Regression	J	0.63	0.97	0.76
		P	0.65	0.09	0.16
	Random Forest	J	0.61	0.72	0.66
		P	0.38	0.28	0.32
	Linear SVC	J	0.63	0.97	0.76
		P	0.66	0.08	0.15
	BERT	J	0.54	0.10	0.18
		P	0.59	0.90	0.70

Table 18: Precision, Recall and F1 Score for MBTI Dimensions on models trained on Reddit data and Twitter as test data

Dimension	Model	Class	Precision	Recall	F1 Score
E/I	Logistic Regression	E	0.36	0.35	0.36
		I	0.60	0.61	0.61
	Random Forest	E	0.38	0.01	0.02
		I	0.60	0.99	0.77
	Linear SVC	E	0.36	0.35	0.36
		I	0.62	0.62	0.62
	BERT	E	0.42	0.00	0.00
		I	0.63	1.00	0.77
S/N	Logistic Regression	N	0.81	0.81	0.81
		S	0.22	0.21	0.21
	Random Forest	N	0.80	1.00	0.89
		S	0.00	0.00	0.00
	Linear SVC	N	0.81	0.84	0.82
		S	0.22	0.19	0.20
	BERT	N	0.78	0.94	0.85
		S	0.24	0.06	0.10
T/F	Logistic Regression	F	0.70	0.54	0.61
		T	0.34	0.50	0.41
	Random Forest	F	0.00	0.00	0.00
		T	0.32	1.00	0.49
	Linear SVC	F	0.69	0.54	0.61
		T	0.34	0.48	0.40
	BERT	F	0.20	0.10	0.12
		T	0.60	0.95	0.72
J/P	Logistic Regression	J	0.56	0.55	0.55
		P	0.47	0.49	0.48
	Random Forest	J	0.55	0.03	0.05
		P	0.46	0.97	0.63
	Linear SVC	J	0.55	0.59	0.57
		P	0.47	0.42	0.44
	BERT	J	0.59	0.97	0.73
		P	0.43	0.03	0.06