THE EFFECTIVENESS OF THE BIG BIRD TRANSFORMER IN TEXT CLASSIFICATION USING SUPPORT VECTOR MACHINES

BRAM DIESCH

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

2044830

COMMITTEE

Dr. Chris Emmery
Bosong Ding (MSc)

LOCATION

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

DATE

June 24th, 2023

WORD COUNT

8569

ACKNOWLEDGMENTS

This thesis was the finalizing part of my masters in Data Science and Society. During the research process I had the chance to delve deeper into the field of Natural Language Processing, an interesting and insightful experience. During the process I had the chance to combine multiple aspects from previous courses, such as Machine Learning, Data Processing and Statistics, the knowledge of which contributed significantly to the research process. I would like to thank Dr. Chris Emmery for his supervision during the thesis process, who helped me with valuable insights and feedback during the process. Happy reading!

THE EFFECTIVENESS OF THE BIG BIRD TRANSFORMER IN TEXT CLASSIFICATION USING SUPPORT VECTOR MACHINES

BRAM DIESCH

Abstract

A key part of understanding one's customers is to understand their personalities, as it allows organizations to develop distinct customer strategies to increase retention and engagement, which can be done through author profiling. While inferring aspects such as age and gender often involve supervised machine learning models, inferring personality may require different methods such as the usage of transformers. However, researchers found that methods such as Support Vector Machines (SVM) versus transformers interchangeably outperform one another. This study aimed to investigate the performance of the Big Bird transformer in combination with a SVM, against those using Tf-Idf vectors. It did so by attempting to predict MBTI-personality using data from Reddit. Multiple SVMs were trained using different kernels (linear, polynomial and rbf). The models' performance was evaluated using 5-fold cross validation, which indicated sufficient model robustness. Model performance varied considerably between models. As the SVM using Tf-Idf achieved accuracy scores ranging from 0.80 to 0.90. While the SVM using Big Bird's embeddings found accuracy scores from 0.65 to 0.74. Clearly indicating the SVM using Tf-Idf offers favorable performance. Similar results were found for metrics such as Recall, Precision and the F1-score, with Big Bird underperforming in all experiments. Best results were found using the polynomial and linear kernels, for the SVMs using Tf-Idf and Big Bird respectively. Big Bird's relatively worse performance could be attributed to the preprocessing of the dataset. As the preprocessing steps included stopword removal and lemmatization, which favors Tf-Idf vectorization. These processes strip sequences of their initial meaning and cohesion, making it more difficult for transformers such as Big Bird to process. Future research should take these steps into consideration, providing more sophisticated methods in handling data for transformers.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

The Reddit MBTI-posts dataset has been acquired from the Emmery et al. (2024) through an in person request during a meeting. The obtained data is anonymized. Work on this thesis did not involve collecting data from human participants or animals. 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 owner was informed about the use of this data for this thesis. All the figures belong to the author. The thesis code can be accessed through Google drive via: Link. Part of the preprocessing, processing, training and analysis code has been written by the author using Phind. Respective libraries used in this study are referenced in the methods section. The reused/adapted code fragments are clearly indicated in the notebook if applicable. In terms of writing, the author used assistance with the language of the paper, the processor in Latex Overleaf was used to improve the author's original content, for paraphrasing, spell checking and grammar. No other typesetting tools or services were used.

2 INTRODUCTION

In today's day and age, audiences, customers and other stakeholders are more segmented than ever before (Khavul et al., 2010). In the literal sense as organizations have turned global, creating a global audience in the process, but also in terms of other demographics like cultural background, political beliefs and personality (Hofstede et al., 2005). These demographics have a significant impact on how organizations must appeal to these audiences as they may differ in their beliefs, values, expectations and interests (Hofstede et al., 2005; Russell & Pratt, 1980). Additionally, it has become more difficult for organizations to gain a clear understanding of their audience, their wants and needs (Ebiaredoh-Mienye et al., 2021). Consequently, this paved the way to new metrics of analysis, such as data science, to analyze audience behavior and defer a clear understanding of their audience composition (Shirazi & Mohammadi, 2019). A key part of understanding one's customers is to understand their personalities, as it allows organizations to develop distinct customer strategies, specifically appealing to increase retention and engagement (Ebiaredoh-Mienye et al., 2021; Esenogho et al., 2022). Studies have shown that a person's personality can be inferred from their written text with the use of natural language processing (dos Santos et al., 2017; Y. Liu et al., 2019; Wu et al., 2020). This process is called author profiling, which aims to infer the author's demographic traits based on their written text (Argamon et al., 2009). While inferring aspects such as age and gender often involve supervised machine

learning models, inferring personality may require different methods such as the usage of transformers (Lanza-Cruz et al., 2024; Santos & Paraboni, 2022). Santos and Paraboni (2022) successfully trained a transformer model on the personality classification tasks by using the BERT-based model. However, they state that other transformer-based machine learning models such as RoBERTa (Y. Liu et al., 2019), ELMo (Peters et al., 2017) or Big Bird (Zaheer et al., 2020) remain untested for this classification task.

Additionally, when applying these models to larger sequences, e.g. larger bodies of text, models like BERT become less optimal due to their self-attention mechanisms (J. Lin et al., 2003). Consequently, Zaheer et al. (2020) proposed that Big Bird, which operates similarly to BERT, is more suitable as it is optimized for longer sequences. However, research applying transformers to this classification problem remains generally limited (Santos & Paraboni, 2022). The limited studies that have incorporated transformers either received favorable result as to these transformers, or prove that already well-established models such as Support Vector Machines (SVM) still remain a potentially superior alternative (Celli & Lepri, 2018; Katna et al., 2022; Wahba et al., 2022). The performance of these SVMs seems to rely on the vectorization method used, with multiple authors reporting conventional vectorizers like Tf-Idf perfroming well against transfromers (Katna et al., 2022; Wahba et al., 2022). However, vectorizers fail to capture the semantic meaning in sequences like transformers do (Santos & Paraboni, 2022). And as Big Bird seems more optimized in dealing with longer sequences (Zaheer et al., 2020), it remains questionable why it has not been compared to other models before. Consequently, it is interesting to determine whether Big Bird could actually compete in this classification task as well.

Moreover, a study by Kalcheva et al. (2020) concluded that the kernel used to tune the SVM is highly dependent on the context and dataset used for the task. Yet, research incorporating the SVM either provides varying results with regards to the most optimal kernel for this task (Celli & Lepri, 2018; Wahba et al., 2022), or fail to provide any specifications regarding this parameter at all (Katna et al., 2022; Wang et al., 2021). It therefore also remains unclear what kernel would fit this specific task best. Consequently, this study aims to contribute to literature by applying Big Bird's embeddings to SVMs in classifying an author's personality traits. Additionally, the study aims to clarify the effectiveness of the Tf-Idf vectorizer against Big Bird and to what extent different kernels can influence the performance. For this purpose the following research question and sub questions are formulated:

Can the Big Bird-transformer improve the model performance of Support Vector Machines in personality prediction through text classification of Reddit posts?

SQ1 *Does the use of Big Bird's embeddings affect the optimal kernel in the SVM?*

SQ2 *How do vectors generated by Tf-Idf affect the SVM's performance compared to Big Bird's embeddings?*

As the Big Bird transformer is generally considered an improvement over the BERT-model (Zaheer et al., 2020), comparing accuracy to non-transformer-based models not only tests its performance on the classification task, but also allows determining whether the Big Bird transformer could become a feasible competitor to the BERT-transformer or SVM in author profiling. Offering researchers and practitioners additional insights into the possible models to use when faced with similar machine learning problems. And extending research to the use of Big Bird in text classification in combination with SVMs.

This introduction is followed by the literature review, which delved deeper into the task of text classification, and highlights the current knowledge regarding the use of transformers and alternative models such as the SVM. Following the literature review, the method section is presented, describing the research approach and specific methods and libraries used to perform the analysis. A preliminary analysis of the initial models was also performed. Following the methods the model, results are presented and discussed. From the discussion, it can be concluded that in this particular research setting, the SVM using Tf-Idf vectors performs better and potentially remains a favorable choice when deciding on which model to use in text classification. However, it remains possible that the handling of the embeddings have skewered the results. Consequently, further research is advised, taking sufficient time to generate the embeddings using the Big Bird transformer. Yet, this research contributes to literature in further establishing the SVM as a suitable candidate in text classification, specifically author profiling. Practitioners are advised to take SVMs into account when dealing with similar machine learning problems.

3 LITERATURE REVIEW

Developments in technology have increased the amount of textual data that is readily available, such as the increased use and applicability of social media, increased cloud storage and other digital tools (da Costa et al., 2023). Many techniques have been developed to deal with these sets of data, such as filtering out spam (W. Liu & Wang, 2010), sentiment analysis (Dawei

et al., 2021), author profiling (B. B. C. da Silva & Paraboni, 2018), and so on. These are prevalent processes that were developed within the field of Natural Language Processing, which revolves around developing and analyzing methods that allow computers to interact with users in human language, such as English (Sharp, 2015). The field of Natural Language Processing thus also pertains to the use of specific tasks in utilizing text data, such as text classification, which is the process of categorizing (part of) a text into a specific category (Wu et al., 2022).

The applicability of Natural Language Processing can thus be considered as a broad spectrum, which can touch upon many use-cases. For example when considering the task of text classification, both spam detection (W. Liu & Wang, 2010) and author profiling (B. B. C. da Silva & Paraboni, 2018) can be considered applications of this particular task, yet potentially totally different use-cases. With spam detection basically pertaining to classifying texts to either the spam category or not (W. Liu & Wang, 2010). Author profiling may be more complex and therefore is somewhat defined differently among authors. Rangel et al. (2015) state author profiling refers to distinguishing between different classes of authors from a psychological viewpoint, by studying their language. B. B. C. da Silva and Paraboni (2018) relate to author profiling as the task of inferring an author's demographic information through the analysis of their written text. Both in essence aim to gather the same information, such as age, gender and personality type (B. B. C. da Silva & Paraboni, 2018; Rangel et al., 2015). Interestingly, some authors even view the task of author profiling to exclusively be related to gather information about an author's personality (Katna et al., 2022), mainly focusing on personality prediction. Author profiling can therefore be considered as a classification task as well, as one attempts to detect certain demographic aspects from an author, with the demographics often consisting of predetermined classes (B. B. C. da Silva & Paraboni, 2018; dos Santos et al., 2017; Y. Liu et al., 2019; Wu et al., 2020).

Initially these demographics often consisted of either age or gender, with the predetermined classes being the age groups and male or female respectively (Deutsch & Paraboni, 2023; Rangel et al., 2013). However, later author profiling extended to more in-depth demographics, such as political alignment(Garcia-Diaz et al., 2022; Rangel et al., 2015), education level(Ashraf et al., 2020) and personality prediction(Santos & Paraboni, 2022). While demographics such as age or gender are useful to know, more complex demographics such as personality could be more useful (Munaf et al., 2009). As for example knowing ones personality gives businesses an indication whether that person will like a certain product or not (Rangel et al., 2015). Alternatively, the effectiveness of certain marketing techniques highly depends on the personality of the receiver (Ebiaredoh-Mienye et al.,

2021; Esenogho et al., 2022). This indicates the importance to move to more complex demographics, and need to develop ways to predict these more accurately.

3.1 *Text classification methods*

When considering text classification tasks, specifically those pertaining to author profiling, many different models have been used (Bonaccorso, 2018; Santos & Paraboni, 2022). Notable models are logistic regression (Gjurković & Šnajder, 2018; Katna et al., 2022; Plank & Hovy, 2015; Wu et al., 2020), decision trees (Bonaccorso, 2018), naive bayes (Bonaccorso, 2018; Katna et al., 2022) and support vector machines (SVM) (Gjurković & Šnajder, 2018; Katna et al., 2022; Rangel et al., 2015; Verhoeven et al., 2016). These models were and still are prevalent in text classification, and in extension author profiling. Interestingly, studies pertaining to author profiling, often apply these models to predicting demographics such as age and gender (Katna et al., 2022; Santos & Paraboni, 2022; Wahba et al., 2022), with fewer applying them for more complex demographics such as personality prediction (Katna et al., 2022; Rangel et al., 2013; Wahba et al., 2022). When studies do aim to predict demographics such as personality, they often do so using numerical features (Utami et al., 2021). This may be due to more complex tasks, such as personality prediction, requiring more advanced machine learning methods, that are better suited for the task (Cervantes et al., 2008; Santos & Paraboni, 2022). A potential alternative may be the use of transformers, which are deep learning architectures that are based on a multi-head attention mechanism, aiming to make the processing of sequential data like text more efficient (T. Lin et al., 2022; Vaswani et al., 2017). Over time, multiple of these transformers have been developed, initially starting out with the Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018). With more transformers based on BERT being created after it's success. Such as ELMo (Peters et al., 2017), RoBERTa (Y. Liu et al., 2019) and Big Bird (Zaheer et al., 2020). These transformers differ from conventional models such as the SVM, partially because they come pre-trained (Vaswani et al., 2017), since they have been trained on data outside of the provided dataset. While for example an SVM is only trained on the data that is fed to it during the training process (Srivastava & Bhambhu, 2010). Additionally, models like the SVM are originally developed to handle data other than specifically sequential or textual data (Srivastava & Bhambhu, 2010), with transformers on the other hand specifically being developed for textual data (Wahba et al., 2022). Specifically because a transformer is meant to understand the context and semantics from within the text (Santos & Paraboni, 2022),

potentially making them more suitable for text classification. Considering transformers such as Big Bird, they are more suited in dealing with longer sequences, which are more common when considering textual data (Zaheer et al., 2020). In comparison, models such as the SVM originally become increasingly computationally expensive when handling greater amounts of data, such as large bodies of text (Cervantes et al., 2008). Additionally, these models often tend to use static word embeddings or bag-of-words in order to perform classification tasks, which fails to capture the semantic meaning and interpretation of the text (Santos & Paraboni, 2022). Consequently, one could argue transformers are a more suitable method in dealing with text classification, specifically personality prediction.

3.2 *Transformers*

The potential use-case of transformers has not gone unnoticed, with an increasing amount of studies shifting their attention to the potential of transformers (Ai et al., 2023; Santos & Paraboni, 2022), and testing predominantly the BERT transformer with increased success. Consequently leading to the creation of the aforementioned transformers like ELMo (Peters et al., 2017), RoBERTa (Y. Liu et al., 2019), and Big Bird (Zaheer et al., 2020). However, despite their potential in text classification, specifically personality prediction, the amount of studies incorporating the transformers remains limited. Some authors have studied the application of transformers in personality prediction (S. C. da Silva & Paraboni, 2023; Humayun et al., 2023; Santos & Paraboni, 2022), but doing so only using BERT. Transformers like ELMO, RoBERTa and Big Bird still remain relatively unknown. These authors, too, highlight the increased performance transformers can offer, and recommend researchers to start incorporating these other transformers as well (Santos & Paraboni, 2022). This recommendation is not unwarranted, as newer transformers are specifically developed to improve upon BERT, such as Big Bird (Zaheer et al., 2020). Big Bird in particular is developed to deal with longer sequences, which is a common occurrence within text classification, as textual works are more commonly of an increased size (Zaheer et al., 2020).

3.3 *Support Vector Machines*

While these transformers show great potential, their conclusive superiority over alternative methods within the natural language processing sphere is yet to be proven (Wahba et al., 2022). Previously, methods such as SVMs were lacking when it comes to large datasets, as its size contributes to a more complex dataset, which is difficult for the SVM to process (Cervantes

et al., 2008). Researchers dealt with this issue by either modifying the SVM to handle the data in chunks, or simplifying the training process in order to reduce the computational expensiveness (Cervantes et al., 2008). And alternatively, select just a representative part of the training set in order to train the model (Awad et al., 2004). This allowed researchers to apply the SVM to larger datasets, but with limited success. As these methods either still resulted in highly complex and computationally expensive models, or having the SVM utilize only a small subset of the available data, potentially hurting its validity (Cervantes et al., 2008).

However, later research shows that SVMs actually do have significant potential in performing text classification tasks (Cervantes et al., 2008; Wahba et al., 2022; Wang et al., 2021). Wang et al. (2021) found that upon increasing the size of the text, the SVM kept stable performance and outperformed other conventional models such as Naive Bayes and Logistic Regression. Research by Celli and Lepri (2018) found that personality prediction via SVMs is possible, reporting the SVM to outperform alternative models. Specifically by using an SVM that incorporates the polynomial kernel with an accuracy of 0.75, 0.15 points higher than a linear SVM, among others. Additionally, Katna et al. (2022) found that SVMs using vectors generated through Tf-Idf outperforms other conventional models, but does require a number of steps to train. They found that the SVMs were able to perform with an accuracy of 0.88 and 0.81, which are considerable numbers. This also shows a significant increase over the SVMs' performance by Celli and Lepri (2018). Wahba et al. (2022) also used the Tf-Idf to vectorize words, using n-grams. They then found that a linear SVM has comparable or even better performance over transformer-based models, such as BERT. The SVM achieved accuracy scores of 0.79, 0.98, 0.93 and 0.82 on different samples. However, performance remained relatively close to the transformer-based models, with accuracy often only differing with only 0.02 points (Wahba et al., 2022). Additionally they reported that in the case a transformer-based model outperformed the SVM, it did so only slightly on a larger dataset containing longer sequences (Wahba et al., 2022). And interestingly, the transformer-based models seem to perform better on datasets which do not contain rare words that are uncommon among other sets (Wahba et al., 2022). This further accentuates how close both transformers and SVMs potentially are in their performance. The variation in results could indicate that either the SVM or transformer could perform better, depending on the size and complexity of the dataset (Celli & Lepri, 2018; Katna et al., 2022; Wahba et al., 2022; Wang et al., 2021).

Moreover, different kernels were used to gain the most optimal performance, Celli and Lepri (2018) reportedly using polynomial, versus Wahba et al. (2022) using a linear kernel. Which constitute to either there being a

non-linear or linear distribution of the data. Sadly, other studies applying the SVM to personality prediction via text classification tasks, often fail to specify these details, leaving room for speculation. However, there is another paper by Kalcheva et al. (2020) aimed to compare different SVM kernels in a text classification task. While they did not disclose any discussion on why certain kernels performed well, they did find that the linear, polynomial and radial base (rbf) kernel are appropriate for this task, with accuracy's above 0.83 being reported (Kalcheva et al., 2020). This is partially consistent with the results of Celli and Lepri (2018) and Wahba et al. (2022), who found optimal performance in the polynomial and linear kernels respectively. Moreover, Kalcheva et al. (2020) concluded that the appropriate kernel is highly dependent on the data and classification task at hand. This conclusion is appropriate given the varying results in optimal kernels found (Celli & Lepri, 2018; Kalcheva et al., 2020; Wahba et al., 2022). Additionally, it is interesting to see whether the use of embeddings in the SVM affects the optimal kernel. This will be investigated further, for which the following subquestions are formulated:

SQ1 *Does the use of Big Bird's embeddings affect the optimal kernel in the SVM?*

What does seem notable is the usage of the Tf-Idf vectorizer, which contributed to the postive performance of the SVM in multiple studies, especially regarding longer sequences (Katna et al., 2022; Wahba et al., 2022). Potentially making it a suitable vectorizer to research further. Considering this, the following sub question has been formulated:

SQ2 *How do vectors generated by Tf-Idf affect the SVM's performance compared to Big Bird's embeddings?*

Regardless of the apparent success of these conventional methods, Cervantes et al. (2008) suggested a different approach is needed. Indicating that a different feature selection method is required, as conventional methods inadequately capture the complexity of larger datasets or sequences. Instead, Cervantes et al. (2008) propose to use Minimal Enclosing Ball clustering in order to gain the appropriate vectors. This could still incorporate the Tf-Idf vectorization to turn the sequences into numerical values, but should improve the utility of the vectors. Cervantes et al. (2008) claims may be warranted, as Wahba et al. (2022) came to a similar conclusion with SVM under performing slightly as the dataset became more complex, suggesting that there could be a point where the dataset becomes too complex that SVMs performance eventually decreases. However, using a complex method in dealing with an already complex dataset may not be wise, as using Minimal Enclosing Ball clustering results in a computationally expensive task (Chan & Pathak, 2014). In combination with Big Bird,

which is aimed for more complex data, the resulting process may prove too expensive. While this topic is interesting and noteworthy, it will therefore not be explored further in this study. As increasing the complexity of the classification task may provide computational expenses that hinder other areas of this study.

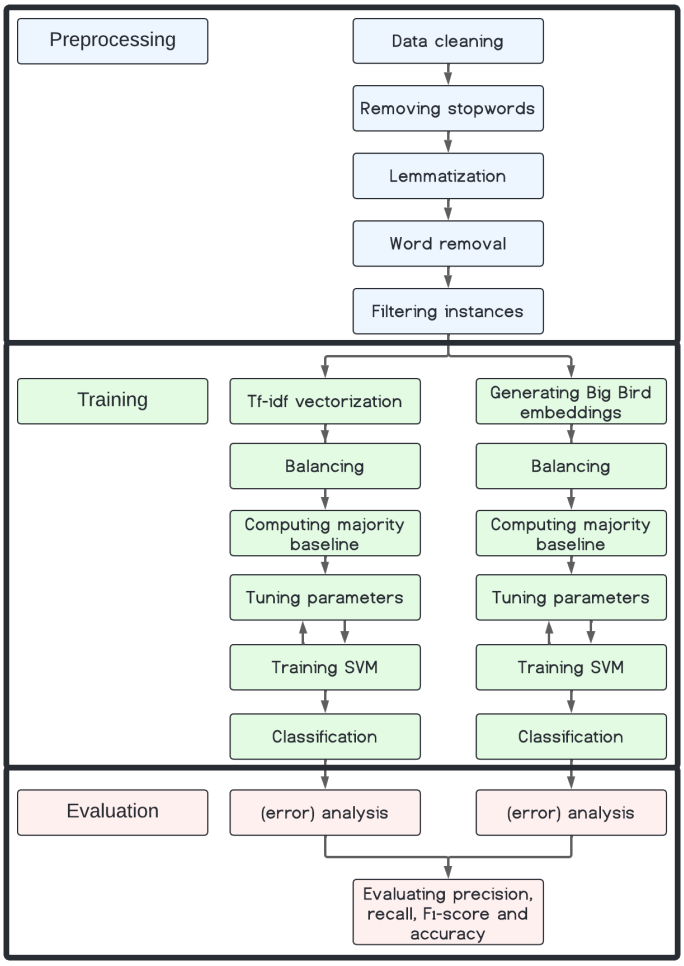
3.4 *Embeddings for SVMs*

Comparing these models to the transformers themselves, developments over the years have thus allowed SVMs to potentially outperform transformers such as BERT, even when considering larger datasets (Wahba et al., 2022). And while transformers still remain promising (Santos & Paraboni, 2022), conventional methods such as SVMs are not off the table. However, even now, Big Bird remains untested and has yet to be compared to these non-transformer alternatives. In light of research regarding the use of SVMs, claims by Zaheer et al. (2020) that Big Bird should be more suited to deal with longer sequences, become more interesting. Not necessarily as a competing model against the SVM, but complementary, using the embeddings generated by the transformer as input for the SVM (Kazameini et al., 2020). The complexity of generating the right word embeddings to input into the SVM (Cervantes et al., 2008), could potentially be negated by Big Birds ability to deal with the large amounts of text with ease (Zaheer et al., 2020). Additionally, as transformers are more capable in dealing with the semantic meaning of the text (Santos & Paraboni, 2022), they may prove to be a useful tool to help increase an SVMs performance.

4 METHOD

In order to address the research questions, both a model with and without embeddings generated through Big Bird were trained. In order to increase the replicability of the study, detailed specifications regarding the research setting, dataset, development steps as well as the methods and libraries used, are specified within this section. With regards to the experimental setting, the study used a Windows 11 machine, containing an AMD Ryzen 9 5900X 12-core processor, as well as 32 Gigabytes of ram. Code development and execution took place using Python 3.11.5 via Visual Studio Code run through an Anaconda environment. The flowchart regarding the research setup is provided in Figure 1. In the sections preprocessing, processing, training and analysis the research steps shown in the figure are discussed.

Figure 1: Flowchart visualizing the research process, from preprocessing up until analysis of the resulting model performance.



4.1 *Personality prediction*

The research will specifically address a text classification problem, focusing on personality prediction in particular. Throughout history multiple measures to classify personality have been developed, such as the Big Five by Fiske (1949) and MBTI by Myers et al. (1985). In terms of performance, Big Five seems more equipped in accurately classifying ones actual personality. However, MBTI still remains as a more popular choice as it is easier to use (Celli & Lepri, 2018). Consequently, the public often chooses to use the MBTI to classify their personality. Additionally, Celli and Lepri (2018) found that the MBTI can be relativelly accurately predicted using the SVM, with lower performance being found in predicting personality based on the Big Five. Because of this, data regarding personality based on MBTI is expected to be more readily available in larger amounts. Moreover, the SVM seems to perform relatively well in predicting MBTI personality flairs (Celli & Lepri, 2018), making MBTI a suitable method to address in testing the effectiveness of Big Bird against SVMs.

4.2 *Dataset description*

The dataset used in this study is one by Emmery et al. (2024) and contains Reddit posts made in the period 2020 to 2022, alongside the authors' respective personality flairs. Authors' MBTI score is based on self-reported MBTI-flairs chosen by authors' to display alongside their name on the subreddit, or inferred by Emmery et al. (2024) when the authors referred to their MBTI score in one of their posts. The personality flairs are consistent with the MBTI personality indicators, which matches the personality scale used for this study. The actual dataset consisted of four smaller subsets, corresponding to each of the MBTI scales: extroversion/introversion, sensing/intuition, perception/judging, feeling/thinking. Each of the subsets initially consisted of three features: *Author_ID*, which identifies which posts are made by the same author; *post*, which contains a string with (part of) an author's Reddit post; and *extroversion*, *sensing*, *judging* and *feeling* respectively, which is a binary scale of whether the author is for example extrovert or not. Most notably, *post* needs to be preprocessed in order to execute the classification task. Since *post* involves the Reddit post or sequence of text the author has shared on the platform. This includes any punctuation, references to an MBTI personality indicator and other unintelligeble words.

Table 1 shows the distribution of the dataset by Emmery et al. (2024). For descriptive purposes, the instances in the raw dataset were also grouped to their respective authors in order to gain an understanding

Table 1: Research dataset descriptives, by authors and instances, dataset from Emmery et al. (2024).

Subset	Authors		Instances	
	Count	Percentage	Count	Percentage
extroversion	438	21.99	9082	22.45
Introversion	1554	78.01	31370	77.55
Sensing	283	14.48	5257	12.56
Intuition	1671	85.52	36582	87.44
Perception	793	40.21	15900	38.44
Judging	1779	59.79	25465	61.56
Feeling	701	33.91	13432	33.92
Thinking	1366	66.09	26168	66.08

of the actual personalities that were represented in the dataset. The distribution of classes is noteworthy, as class imbalance can be inferred from the data. This constitutes to the idea that one's personality plays a role in which social media platform they use and thus influences the distribution of the sample drawn from Reddit (Chen & Peng, 2023).

4.3 *Preprocessing*

The Reddit data by Emmery et al. (2024) contained the raw data, without any processing being done. Consequently requiring preprocessing in order to clean the data and prepare it for the classification task. For both the SVMs with and without Big Bird's embeddings, the preprocessing steps are equal, to ensure the experimental conditions only differ in either the use of Big Bird or Tf-Idf. Data was first loaded using the pandas library (McKinney et al., 2010). First, symbols and other forms of punctuation were removed, using the *re* library by Van Rossum (2020). Next, stopwords were removed, words were lemmatized and non-English words were removed, all using the *nltk* library by Bird et al. (2009). Stopwords were removed in order to improve model performance and reduce the size of the dataset Silva (2003). Lemmatization was used to return words to their base form, while keeping the meaning of the word in its context intact (Arimbawaa & ERa, 2017). Which is beneficial when attempting to interpret the sequences posted by authors. Non-English words were removed to reduce noise in this classification task, as most instances consisted of English words. This also included removing an mentioning of MBTI-flairs as having MBTI personality indicators present in the dataset would introduce unnecessary

noise, and possibly influence its predictions. Finally, instances were filtered from the dataset when they contained less than 5 unique words, as they are considered to either be spam or not contribute enough to this classification task (Baillargeon & Lamontagne, 2024). These steps resulted in a sufficiently clean dataset for this classification task.

4.4 *Processing*

Processing mainly constituted to generating the appropriate embeddings and preparing the data to be interpreted by the SVM. Firstly, in order to generate the embeddings, the *transformers* library version 4.28.1 was used (Inc., 2024b). As well as the *tokenizers* library version 0.13.2 (Inc., 2024a). The embeddings were generated using the pretrained Big Bird transformer by Zaheer et al. (2020) using the big bird roberta base (Inc., 2024b). In order to ensure the embeddings are interpretable to the SVM and feature space is appropriate, embeddings were averaged in order to create a one dimensional array per instance. Generated embeddings were then written to a new column in the dataset called *embeddings*.

For the regular SVM, instead of embeddings, the Tf-Idf vectorizer was used to vectorize the text, imported from the sklearn library (Pedregosa et al., 2011). Similarly to Big Birds embeddings, this task was executed for each of the subsets. However, these vectors were not saved to a separate file, but immediately used as input for the models.

4.5 *Training*

After (pre)processing, training of the models took place, which totalled to 24 models. Twelve for both the Tf-Idf SVM and Big Bird SVM. For each of the subsets, extroversion, sensing, feeling and judging, the SVMs have been trained on three different kernels: linear, polynomial and rbf.

The SVM used was based on the works of Vapnik (1995). The data was randomly split into training and test sets, with 80% belonging to the training, and 20% to the test set. The split was done in an instance level. This was done for each of the subsets separately. For replication purposes, a random state of 42 was used. Following the split, a benchmark for the model was generated using a majority baseline. It was generated using the *DummyClassifier* from the sci-kit learn library (Pedregosa et al., 2011). In order to deal with class imbalance, the majority class of the training set was undersampled, to achieve an equal distribution of the decision classes for the training set. In order to maintain the validity of the sample, the test set remained as is, so it adequately resembled the class distributions within the sample. The undersampling used the *RandomUnderSampler* from

Table 2: Majority baselines. Subsets are referred to as extroversion/ introversion (E/I), sensing/ intuition (S/N), feeling/ thinking (F/T) and judging/ perceiving (J/P) respectively.

	Subset			
Model	E/I	S/N	F/T	J/P
Majority	0.78	0.88	0.67	0.61

the Imbalanced-learn library (Lemaître et al., 2017). After the re-balancing, the actual model was to be trained. For the SVM, a C-Support Vector Machine was used, being imported from the Sklearn library (Pedregosa et al., 2011). Alternatively, for large datasets a Linear C-Support Vector Machine is recommended as it is specifically developed to handle large datasets with more ease (Pedregosa et al., 2011). However, due to the available literature mainly relating to the C-Support Vector Machine, this study also maintained that model for the analysis.

For the SVM, certain parameters could be tuned to increase its performance (Pedregosa et al., 2011). In this case, the C regularization parameter was set to 1 and gamma was set to 'scale' to adequately deal with the comlexity of the vectors/ embeddings (Pedregosa et al., 2011). However, the kernels were tuned, varying between linear, polynomial and rbf, in order to answer the research sub questions. The class_weight was initially set to 'balanced' to automatically deal with the class weights. Different combinations of kernels and class weights were tested and used for input in the analysis. These steps were the same for both the SVM using Big Bird's embeddings and the SVM using Tf-Idf vectors.

4.6 *(preliminary) Evaluation*

Metrics used to evaluate and compare the models' performance were precision, recall, F1-score and accuracy. While accuracy is an important metric to take into account when evaluating model performance, the analysis of the models also incorporated and relied on the other metrics in order to account for the class imbalance, apart from the re-balancing of the training data.

The majority baselines for the analysis are presented in Table 2. They resemble the occurrence of the majority class within each of the subsets and serve as a baseline for the models to achieve. The baselines show the class imbalance present in the testing sets, even after the undersampling procedure. As the test set remains true to the distribution of the sample in order to maintain validity.

Table 3: 5-fold cross validation results for the SVM using Tf-Idf vectors. Cells include the folds' mean and standard deviation (SD). Subsets are referred to as extroversion/ introversion (E/I), sensing/ intuition (S/N), feeling/ thinking (F/T) and judging/ perceiving (J/P) respectively.

	SVM Tf-Idf			
Kernel	E/I	S/N	F/T	J/P
Linear	0.76; SD ±.008	0.82; SD ±.007	0.79; SD ±.004	0.77; SD ±.005
Poly	0.79; SD ±.012	0.79; SD ±.008	0.80; SD ±.003	0.80; SD ±.004
Rbf	0.78; SD ±.011	0.82; SD ±.009	0.80; SD ±.005	0.79; SD ±.004
Baseline	0.78	0.88	0.67	0.61

In order to test the robustness of the trained models, a 5-fold cross validation has been performed. Tables 3 and 4 show results regarding this cross validation for the SVMs using Tf-Idf vectorizer and Big Bird's embeddings respectively. The average accuracy score and standard deviation in accuracy scores within the folds are listed. This gives a clear overview of the robustness of the trained models. From Table 3, the results of the cross validation from the SVM using Tf-Idf can be seen. The standard deviations show values no higher than 0.01, which indicates the variation in the different folds' performance is low, and the models being considered robust.

From Table 4 it is notable that model performance seems relatively consistent, with no standard deviations higher than 0.011 being found. This shows the SVMs using Big Birds embeddings are consistent in their predictions. However, interestingly standard deviation seems to differ slightly between different subsets. With the lowest standard deviations being found for the feeling subset (0.005; 0. 007 and 0.007) for the linear, polynomial and rbf kernels respectively. Compared to the sensing subset that reports standard deviations of 0.012, 0.012, and 0.01 being found. While still acceptable values, the difference between the subsets remains worth noting. Concluding, the models were deemed sufficiently robust to continue the evaluation.

5 RESULTS

This section highlights notable results using multiple tables. The metrics accuracy, precision, recall and the F1-score are discussed separately. Additionally, error analysis, specifically the effects of class imbalance are incorporated within the sections.

Table 4: 5-fold cross validation results for the SVM using Big Bird embeddings. Cells include the folds' mean and standard deviation (SD). Subsets are referred to as extroversion/ introversion (E/I), sensing/ intuition (S/N), feeling/ thinking (F/T) and judging/ perceiving (J/P) respectively.

	SVM Big Bird			
Kernel	E/I	S/N	F/T	J/P
Linear	0.66; SD ±.006	0.73; SD ±.12	0.71; SD ±.005	0.68; SD ±.008
Poly	0.65; SD ±.009	0.72; SD ±.012	0.70; SD ±.006	0.68; SD ±.007
Rbf	0.65; SD ±.007	0.72; SD ±.01	0.70; SD ±.007	0.67; SD ±.007
Baseline	0.78	0.88	0.67	0.61

Table 5 shows the accuracy scores for each of the models. Investigating the SVM using Tf-Idf vectorization, the lowest accuracy found was 0.75 on the extroversion subset, using the linear kernel. The highest accuracy found was on the sensing subset using the polynomial kernel. On average, the SVM using Tf-Idf achieved an accuracy of about 0.84. In all subsets, the Tf-Idf vectorizer was able to manage to provide an accuracy above the majority baseline. Notably, the highest scores for each of the subsets was found using the polynomial kernel, and on average the polynomial kernel achieved higher results compared to other kernels (linear = 0.79; polynomial = 0.84; rbf = 0.81).

Table 5: Accuracy scores for each of the models. The highest found accuracies are highlighted for each fo the subsets. Subsets are referred to as extroversion/ introversion (E/I), sensing/ intuition (S/N), feeling/ thinking (F/T) and judging/ perceiving (J/P) respectively.

	SVM Tf-Idf				SVM Big Bird			
Kernel	E/I	S/N	F/T	J/P	E/I	S/N	F/T	J/P
Linear	0.75	0.83	0.79	0.78	**0.68**	**0.74**	**0.72**	**0.69**
Poly	**0.81**	**0.90**	**0.81**	**0.83**	0.66	0.73	**0.72**	0.68
Rbf	0.78	0.86	0.80	0.80	0.65	**0.74**	0.71	0.68
Majority	0.78	0.88	0.67	0.61	0.78	0.88	0.67	0.61

For the SVM using Big Bird, the lowest accuracy found was 0.65 on the extroversion subset, using the rbf kernel. Interestingly, performance on both the SVM with Tf-Idf and Big Bird achieved the lowest performance on this subset. The highest accuracy found was 0.74 on the sensing subset, both on the model with the linear and the rbf kernel. Similarly to the lowest score, both the SVM with the Tf-Idf and Big Bird achieved the highest performance on the sensing subset. On average, the SVM using

Big Bird's embeddings achieved an accuracy score of 0.70 with the linear kernel performing best over all the subsets (linear = 0.71; polynomial = 0.70; rbf = 0.70). However, the difference per kernel is relatively small, with average accuracy varying only with 0.01 units. Comparing these results to the Tf-Idf, Big Bird seems to underperform in all experiments, with accuracy on average being 0.13 points lower.

5.1 *Precision*

Table 6 shows the models' respective precision values. When comparing the SVM with Tf-Idf vectors, the model shows consistently high precision across all categories, with the highest precision being 0.97 when classifying intuition, with high scores being obtained across all kernels. Introversion also reported high precision values (0.92). The SVM appears capable in accurately identifying classes, but is biased towards majority classes when class imbalance is present.

Table 6: Precision scores for each of the models. Subsets are referred to as extroversion/ introversion (E/I), sensing/ intuition (S/N), feeling/ thinking (F/T) and judging/ perceiving (J/P) respectively.

		Personality classifier							
Model	Kernel	E	I	S	N	F	T	J	P
SVM Tf-Idf	Linear	0.47	0.92	0.41	0.90	0.65	0.89	0.85	0.69
	Poly	0.56	0.92	0.57	0.95	0.87	0.71	0.87	0.76
	Rbf	0.51	0.92	0.46	0.97	0.68	0.89	0.86	0.72
SVM Big Bird	Linear	0.38	0.89	0.29	0.95	0.57	0.84	0.78	0.58
	Poly	0.37	0.88	0.28	0.94	0.56	0.83	0.78	0.57
	Rbf	0.36	0.88	0.28	0.94	0.55	0.83	0.77	0.57

The Big Bird model demonstrates mixed precision outcomes, with the highest precision being 0.89 found when predicting Intuition. The lowest score was found when predicting sensing 0.29. In both subsets extroversion and sensing, a high discrepancy between the two predicted classes can be seen. With more balanced precision values being found in in the feeling and thinking subsets. It seems that the models suffer from a significant amount of false positive predictions. Possibly due to the nature of the subset, as the test set does maintain the class imbalance present in the sample. Notable is however that again, the SVM using Tf-Idf vectors provides better precision results.

When investigating the kernels in particular, the polynomial and rbf kernels show on average slightly better results than the linear kernel, when used by the SVM with Tf-Idf. Remarkably, the SVM using Big Bird's

embeddings shows the linear and rbf kernels provide better results. This shows that depending on the use of Tf-Idf or Big Bird, the most optimal performing kernel seems to vary.

5.2 *Recall*

Recall seems to provide a similar image to precision. As seen in Table 7, recall values for the SVM using Tf-Idf seem consistently higher than its Big Bird counterpart, which is not surprising given the previous metrics also supporting this. With regards to the kernels, all show competitive results. The linear kernel shows relatively high recall values over all the subsets. Similarly, the polynomial and rbf kernels have slightly more variation in the values they return, yet still remaining relatively high. Interestingly, recall seems to be more balanced between classes of imbalanced subsets, compared to the accuracy and precision results. It shows that the models are relatively capable of capturing the true positive predictions among the total set of predictions.

Table 7: Recall scores for each of the models. Subsets are referred to as extroversion/ introversion (E/I), sensing/ intuition (S/N), feeling/ thinking (F/T) and judging/ perceiving (J/P) respectively.

		Personality classifier							
Model	Kernel	E	I	S	N	F	T	J	P
SVM Tf-Idf	Linear	0.76	0.75	0.81	0.84	0.75	0.80	0.78	0.78
	Poly	0.75	0.83	0.67	0.93	0.75	0.84	0.84	0.80
	Rbf	0.78	0.78	0.79	0.87	0.80	0.81	0.81	0.80
SVM Big Bird	Linear	0.71	0.67	0.74	0.74	0.74	0.71	0.68	0.69
	Poly	0.70	0.65	0.70	0.74	0.72	0.72	0.67	0.70
	Rbf	0.70	0.64	0.69	0.74	0.73	0.70	0.67	0.69

5.3 *F1-score*

With regards to the SVM with Tf-Idf, relatively high F1-scores can be found consistently, as seen in table 8. Most notable is that F1-scores pertaining to the subsets feeling and judging have only a slight variation in F1-score for each of the classes. This indicates that the models seem to perform relatively well in both identifying the positive and negative classes within the subset. However, more variation can be identified when investigating the extroversion and sensing subsets. For example, the SVM using a linear kernel reports an F1-score of 0.58 and 0.83 for the minority and majority

class respectively. Moreover, the sensing subset reports F1-scores of 0.54 and 0.90 for the minority and majority classes, an even greater difference than extroversion. While the class imbalance in these subsets is apparent, it is interesting that the F1-scores still remain relatively acceptable. The more balanced subsets, feeling and judging, also show differences between classes, however this difference is considerably lower.

With regards to the SVM using Big Bird's embeddings, similar conclusions can be drawn. F1-scores for the subsets show relative consistency between both the majority and minority classes, with still a slight decrease in F1-score being observed when considering the minority class. Similarly to the SVM using Tf-Idf, the extroversion and sensing subsets show an increased variation between the majority and minority classes, with in this case the minority class performing poorly compared to the majority class. The lowest score was 0.40 for the imbalanced subsets, with 0.62 being found for relatively balanced subset. This is a considerable difference, accentuating the class imbalance heavily impacts the predictive performance of the models. High scores were not that different, with 0.76 and 0.84 being found for the imbalanced subsets extroversion and sensing respectively, and 0.77 and 0.73 being found for the relatively balanced subsets feeling and judging. Indicating that the class imbalance has a considerable effect on the effectiveness of Big Bird's model. With these results it becomes clear that the SVM using Tf-Idf outperforms the SVM using Big Bird's embeddings on all accounts.

Table 8: F1-scores for each of the models. Subsets are referred to as extroversion/ introversion (E/I), sensing/ intuition (S/N), feeling/ thinking (F/T) and judging/ perceiving (J/P) respectively.

		Personality classifier							
Model	Kernel	E	I	S	N	F	T	J	P
SVM Tf-Idf	Linear	0.58	0.83	0.54	0.90	0.72	0.83	0.81	0.73
	Poly	0.64	0.87	0.61	0.94	0.73	0.86	0.86	0.78
	Rbf	0.61	0.85	0.58	0.92	0.73	0.85	0.83	0.76
SVM Big Bird	Linear	0.50	0.76	0.42	0.84	0.63	0.77	0.73	0.63
	Poly	0.48	0.75	0.40	0.83	0.63	0.77	0.72	0.63
	Rbf	0.48	0.74	0.40	0.83	0.63	0.76	0.72	0.62

Kernels show a variation in performance, which in this case seems caused by the imbalanced subsets. However, all kernels perform relatively well. Specifically for the SVM using Tf-Idf, the polynomial kernel provides the best results. With regards to the SVM using Big Bird's embeddings, similar performance across all kernels can be spotted. However, the linear kernel seems to perform slightly better, providing an F1-score 0.01 or 0.02

points higher than the polynomial or rbf kernel. This is consistent with earlier metrics, where the linear kernel also seemed to slightly outperform other kernels.

6 DISCUSSION

In this study, the Big Bird transformer was evaluated against regular vectorization methods like the Tf-Idf, and applied to an SVM. In order to test whether Big Bird's potential superior performance in text classification would hold out against a more conventional SVMs. Moreover, parameters such as the use of different kernels were applied in order to investigate whether they affect the effectiveness of either Big Bird's embeddings or Tf-Idf's vectors. As well as whether these embeddings or vectors require different kernels to optimize performance. Therefore, this study aimed to address the following questions:

> *Can the Big Bird-transformer improve the model performance of Support Vector Machines in personality prediction through text classification of Reddit posts?*

SQ1 *Does the use of Big Bird's embeddings affect the optimal kernel in the SVM?*

SQ2 *How do vectors generated by Tf-Idf affect the SVM's performance compared to Big Bird's embeddings?*

Regarding the effectiveness of Big Bird's embeddings against Tf-Idf's vectors, it was expected for Big Bird to increase model performance. In line with the findings of Santos and Paraboni (2022), Big Bird's embeddings should be able to capture the semantic meaning of the sequence or text, which Tf-Idf's vectors are less capable of. When addressing longer sequences, being able to capture the authors' meaning could prove useful in text classification. Extending to personality prediction in this case. Regardless, Tf-Idf's vectors have consistently been proved effective in text classification and personality prediction (Katna et al., 2022; Wahba et al., 2022), establishing it as a valid competitor against Big Bird's performance. Moreover, SVM's can be tuned using multiple parameters, one of which is the kernel (Pedregosa et al., 2011; Vapnik, 1995). However, studies either lack enclosing which kernel provided better performance, or reported varying kernels being the best choice Celli and Lepri (2018) and Wahba et al. (2022). This was in line with the works of (Kalcheva et al., 2020), who concluded the choice of kernel is highly dependent on the data and research setting. However, based on the presented studies, it was expected that either the linear or polynomial kernel would perform best, with the rbf being a potential third candidate.

6.1 *SVM performance*

In the results section the trained models' performance was evaluated using accuracy, precision, recall and the F1-score. This provided interesting insights. Firstly, when comparing the SVM using Tf-Idf vectors, compared to the SVM using Big Bird's embeddings, the SVM with Tf-Idf seems to perform better overall. Across all metrics, the SVM with Tf-Idf was able to achieve higher scores. This is in line with the findings of Wahba et al. (2022), who showed comparable or even increased performance of SVM using Tf-Idf over transformer-based models. Furthermore, scores from this study are consistent with scores from previous studies. Celli and Lepri (2018) reported an accuracy of 0.75, for their SVM without the use of Tf-Idf vectors. Interestingly, higher accuracies were reported for SVMs with the use of Tf-Idf vectors, with Wahba et al. (2022) reporting accuracies varying from 0.78 to 0.98, and Katna et al. (2022) reporting accuracy 0.81 and 0.88. It seems our SVM with Tf-Idf vectors performs relatively similar to these other studies incorporating the Tf-Idf, with our models showing varying accuracies from 0.81 to 0.90. Indicating the use of Tf-Idf improves the models' performance.

The positive performance of the SVM using Tf-Idf can have multiple reasons. The Tf-Idf vectorizer is a robust method that has been successfully used in many studies (Katna et al., 2022; Wahba et al., 2022). This is consistent with this study's results, as the 5-fold cross-validation showed reasonable performance across all folds with minimal deviation in scores. Additionally, the fact that it already is a vectorizer for text based data (Pedregosa et al., 2011), it would be reasonable to assume it would perform adequately in performing binary text classification (Celli & Lepri, 2018; Katna et al., 2022; Wahba et al., 2022; Wang et al., 2021).

Relating to the SVM using Big Bird's embeddings, it is interesting to explore why it underperforms the Tf-Idf. Firstly, Big Bird's performance is said to primarily come from its ability to deal with longer sequences of text (Zaheer et al., 2020). Moreover, a transformer is especially meaningful when able to capture the semantic meaning of the text (Santos & Paraboni, 2022). However, the preprocessing included a number of steps that for one, reduced words to a more basic form. For this step, a lemmatizer was used to ensure the word still retains some of its meaning in the text (Arimbawaa & ERa, 2017). However, this may have had an impact on Big Bird's ability to fully capture the meaning of the respective words in relation to the rest of the text (Bass et al., 2019). Additionally, stopwords and words that were not specifically in the English dictionary were removed. While stopwords may not be that relevant in the text, they do tie different words together to build the narrative of the text (Bass et al., 2019). Bass et al.

(2019) found that both lemmatization and stopword removal resulted in lower model performance for transformer-based models. They go on to state that these two processes affect transformers' ability to understand the context among words. This is a problem the Tf-Idf vectorizer does not have, as it does not incorporate the words' context (Ramos et al., 2003). Additionally, in order to generate the embeddings in a reasonable way, the embeddings were averaged to reduce the dimensionality during the computation process. While this allowed the embeddings to be used by the SVM, this may have had an effect on the embeddings' effectiveness. As the averaged embeddings now no longer represented the relationship among specific words.

6.2 *Kernels*

The research process also included the tuning of specific kernels, namely linear, polynomial and rbf. Based on previous studies incorporating results regarding the kernels used (Celli & Lepri, 2018; Wahba et al., 2022), it was expected that either the linear kernel and/ or the polynomial kernel would perform best within this classification task. Kalcheva et al. (2020) also addressed the linear and polynomial kernels being appropriate parameters for this task, alongside the rbf. Depending on the evaluation metric clear differences in optimal performance for the SVM with Tf-Idf's vectors, compared to the SVM with Big Bird's embeddings were found. With accuracy scores being highest over all the subsets on the polynomial kernel for the SVM with Tf-Idf vectors. And the linear kernel performing best over all the subsets for the SVM with Big Bird's embeddings. For the SVM using Tf-Idf vectors, the polynomial kernel performed clearly better than the linear and rbf kernels. However, the SVM using Big Bird's embeddings, the linear kernel only slightly outperformed the alternatives. This is interesting and shows that indeed the optimal type of kernel used may be partially dependent on the use of either Tf-Idf, or Big Bird (Kalcheva et al., 2020). This could indicate that the embeddings and decision variable are linearly separated. With the polynomial kernel indicating that it is non-linear when dealing with Tf-Idf vectors (Pedregosa et al., 2011).

6.3 *Limitations*

This study had a number of limitations. First and foremost, Big Bird's embeddings were generated in such a way that could likely have diminished their effectiveness. This was due to the limited time horizon of the research project. With Big Bird's embeddings underperforming against the SVM using Tf-Idf vectors, it seems reasonable that these adjustments

to Big Bird have had an effect. It is therefore recommended that future studies take ample more time to develop their data and models, to ensure the embeddings are generated appropriately. Moreover, the preprocessing steps included lemmatization and stopword removal, which favors the Tf-Idf, that would not be able to capture relationships among words (Ramos et al., 2003). However, Big Bird's performance would have been affected, as lemmatization and stopword removal hinders the transformer's ability to interpret the sequences in their most natural form (Bass et al., 2019). Future research would benefit from having two separate preprocessing steps being done. While this increases the differences in the research setting outside of the main models (SVM with Tf-Idf or SVM with Big Bird), this would allow the datasets to be processed in a way that suits both the vectorizer and transformer. With regards to the training of the models, the train-test split could have skewed the results. Authors could have made multiple posts on Reddit, which were then stored in multiple instances. As the train-test split occured on an instance basis, the same author could appear in both the train and test set, possibly allowing the model to identify the author with more ease and thus correctly predicting their personality. Future research is adviced to instead apply the train-test split on an author basis, ensuring the same author cannot occur in both sets. Additionally, the tuning process regarding the SVM's was limited. With only the linear, polynomial and rbf kernels being adjusted. Future research would benefit from incorporating more tuning parameters in their study to examine the full potential of the SVM.

Apart from the limitations regarding the research process, the dataset also is subject to limitations. Most importantly the dataset consists of Reddit posts, which is a particular online platform. As described by Chen and Peng (2023), one's personality plays a role in which social media platform one uses. It is therefore likely that Reddit users have a different personality from for example Twitter users (Chen & Peng, 2023). Consequently, the results from this study may not be entirely generalizable to other platforms, as the distribution of personality types may appear differently from this study. Future research is advised to take this into consideration, and extend this research to other platforms to test its generalizability.

6.4 *Contributions*

Based on the study, the use of the SVM with Tf-Idf vectors remains a remarkably useful model when attempting to perform text classification. Practitioners are therefore recommended to consider its possibilities within their projects. However, Big Bird's transformer could still prove useful if

the data is prepared more fittingly to the transformer, and more time is taken to train the model properly. This research contributes to literature by shedding more light on the use of transformers within text classification, and provides more insights into the use of Big Bird. Additionally, this study contributes to literature by providing more insights into the application of SVM and Big Bird in author profiling.

7 CONCLUSION

Returning to the main research question:

> *Can the Big Bird-transformer improve the model performance of Support Vector Machines in personality prediction through text classification of Reddit posts?*

This study found that the Big Bird transformer does not improve the model performance of Support Vector Machines in personality through text classification. As mentioned in the discussion, it is likely that the processing of the embeddings and preprocessing of the dataset played a role in Big Bird's relatively low effectiveness. Future researchers are advised to take care when taking similar steps in their research process to prevent this issue from arising.

REFERENCES

Ai, Z., Yijia, Z., & Mingyu, L. (2023). A domain knowledge transformer model for occupation profiling. *International Journal of Computational Intelligence Systems*, *16*(1), 1–13.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, *52*(2), 119–123.

Arimbawaa, I. G. A. P., & ERa, N. A. S. (2017). Lemmatization in balinese language. *Jurnal Elektronik Ilmu Komputer Udayana p-ISSN*, *2301*, 5373.

Ashraf, M. A., Nawab, R. M. A., & Nie, F. (2020). Author profiling on bilingual tweets. *JOURNAL OF INTELLIGENT & FUZZY SYSTEMS*, *39*(2), 2379–2389. https://doi.org/10.3233/JIFS-179898

Awad, M., Khan, L., Bastani, F., & Yen, I.-L. (2004). An effective support vector machines (svms) performance using hierarchical clustering. *16th IEEE International Conference on Tools with Artificial Intelligence*, 663–667. https://doi.org/10.1109/ICTAI.2004.26

Baillargeon, J.-T., & Lamontagne, L. (2024). Assessing the impact of sequence length learning on classification tasks for transformer encoder models. *The International FLAIRS Conference Proceedings*, *37*.

Bass, C. R., Benefield, B., Horn, D., & Morones, R. (2019). Increasing robustness in long text classifications using background corpus knowledge for token selection. *SMU Data Science Review*, *2*(3), 10.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Bonaccorso, G. (2018). *Machine learning algorithms: Popular algorithms for data science and machine learning*. Packt Publishing Ltd.

Celli, F., & Lepri, B. (2018, January). Is big five better than mbti? https://doi.org/10.4000/books.aaccademia.3147

Cervantes, J., Li, X., Yu, W., & Li, K. (2008). Support vector machine classification for large data sets via minimum enclosing ball clustering. *Neurocomputing*, *71*(4-6), 611–619.

Chan, T. M., & Pathak, V. (2014). Streaming and dynamic algorithms for minimum enclosing balls in high dimensions. *Computational Geometry*, *47*(2), 240–247.

Chen, M., & Peng, A. Y. (2023). Why do people choose different social media platforms? linking use motives with social media affordances and personalities. *Social Science Computer Review*, *41*(2), 330–352.

da Costa, L. S., Oliveira, I. L., & Fileto, R. (2023). Text classification using embeddings: A survey. *KNOWLEDGE AND INFORMATION SYSTEMS*, *65*(7), 2761–2803. https://doi.org/10.1007/s10115-023-01856-z

da Silva, B. B. C., & Paraboni, I. (2018). Personality recognition from facebook text. *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, 107–114.

da Silva, S. C., & Paraboni, I. (2023). Politically-oriented information inference from text. *Journal of Universal Computer Science*, *29*(6), 569.

Dawei, W., Alfred, R., Obit, J. H., & On, C. K. (2021). A literature review on text classification and sentiment analysis approaches. In R. Alfred, H. Iida, H. Haviluddin, & P. Anthony (Eds.), *Computational science and technology* (pp. 305–323). Springer Singapore.

Deutsch, C., & Paraboni, I. (2023). Authorship attribution using author profiling classifiers. *Natural Language Engineering*, *29*(1), 110–137. https://doi.org/10.1017/S1351324921000383

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

dos Santos, V. G., Paraboni, I., & Silva, B. B. C. (2017). Big five personality recognition from multiple text genres. *Text, Speech, and Dialogue: 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings 20*, 29–37.

Ebiaredoh-Mienye, S. A., Esenogho, E., & Swart, T. G. (2021). Artificial neural network technique for improving prediction of credit card default: A stacked sparse autoencoder approach. *International Journal of Electrical and Computer Engineering*, *11*(5), 4392.

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*.

Esenogho, E., Mienye, I. D., Swart, T. G., Aruleba, K., & Obaido, G. (2022). A neural network ensemble with feature engineering for improved credit card fraud detection. *IEEE Access*, *10*, 16400–16407.

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *The Journal of Abnormal and Social Psychology*, *44*(3), 329.

Garcia-Diaz, J. A., Jimenez-Zafra, S. M., Martin Valdivia, M.-T., Garcia-Sanchez, F., Urena-Lopez, L. A., & Valencia-Garcia, R. (2022). Overview of politices 2022: Spanish author profiling for political ideology. *PROCESAMIENTO DEL LENGUAJE NATURAL*, (69), 265–272. https://doi.org/10.26342/2022-69-23

Gjurković, M., & Šnajder, J. (2018). Reddit: A gold mine for personality prediction. *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, 87–97.

Hofstede, G., Hofstede, G. J., & Minkov, M. (2005). *Cultures and organizations: Software of the mind* (Vol. 2). Mcgraw-hill New York.

Humayun, M. A., Yassin, H., Shuja, J., Alourani, A., & Abas, P. E. (2023). A transformer fine-tuning strategy for text dialect identification. *Neural Computing and Applications*, *35*(8), 6115–6124.

Inc., H. F. (2024a). Tokenizers [Accessed: 2024-05-15]. https://github.com/huggingface/tokenizers

Inc., H. F. (2024b). Transformers [Accessed: 2024-05-15]. https://github.com/huggingface/transformers

Kalcheva, N., Karova, M., & Penev, I. (2020). Comparison of the accuracy of svm kemel functions in text classification. *2020 International Conference on Biomedical Innovations and Applications (BIA)*, 141–145. https://doi.org/10.1109/BIA50171.2020.9244278

Katna, R., Kalsi, K., Gupta, S., Yadav, D., & Yadav, A. K. (2022). Machine learning based approaches for age and gender prediction from

tweets. *MULTIMEDIA TOOLS AND APPLICATIONS*, *81*(19), 27799–27817. https://doi.org/10.1007/s11042-022-12920-1

Kazameini, A., Fatehi, S., Mehta, Y., Eetemadi, S., & Cambria, E. (2020). Personality trait detection using bagged svm over bert word embedding ensembles.

Khavul, S., Peterson, M., Mullens, D., & Rasheed, A. A. (2010). Going global with innovations from emerging economies: Investment in customer support capabilities pays off. *Journal of International Marketing*, *18*(4), 22–42.

Lanza-Cruz, I., Berlanga, R., & Aramburu, M. J. (2024). Multidimensional author profiling for social business intelligence. *Information Systems Frontiers*, *26*(1), 195–215.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, *18*(17), 1–5. http://jmlr.org/papers/v18/16-365.html

Lin, J., Quan, D., Sinha, V., Bakshi, K., Huynh, D., Katz, B., & Karger, D. R. (2003). What makes a good answer? the role of context in question answering. *INTERACT*.

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*, *3*, 111–132. https://doi.org/https://doi.org/10.1016/j.aiopen.2022.10.001

Liu, W., & Wang, T. (2010). Index-based online text classification for sms spam filtering. *J. Comput.*, *5*(6), 844–851.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

McKinney, W., et al. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, *445*, 51–56.

Munaf, S., Nisa, U., Shaheen, A., Hussain, S., & Kamrani, F. (2009). Personality type, gender and age difference: A study of customers brand loyalty. *5*, 38–53.

Myers, I. B., McCaulley, M. H., & Most, R. (1985). Manual: A guide to the development and use of the myers-briggs type indicator. *(No Title)*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peters, M. E., Ammar, W., Bhagavatula, C., & Power, R. (2017). Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

Plank, B., & Hovy, D. (2015). Personality traits on twitter—or—how to get 1,500 personality tests in a week. *Proceedings of the 6th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 92–98.

Ramos, J., et al. (2003). Using tf-idf to determine word relevance in document queries. *Proceedings of the first instructional conference on machine learning*, *242*(1), 29–48.

Rangel, F., Celli, F., Rosso, P., Martin, P., Stein, B., Daelemans, W., et al. (2015). Overview of the 3rd author profiling task at pan 2015. *CLEF2015 Working Notes. Working Notes of CLEF 2015-Conference and Labs of the Evaluation forum.*

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., & Inches, G. (2013). Overview of the author profiling task at pan 2013. *CLEF conference on multilingual and multimodal information access evaluation*, 352–365.

Russell, J. A., & Pratt, G. (1980). A description of the affective quality attributed to environments. *Journal of personality and social psychology*, *38*(2), 311.

Santos, V. G. d., & Paraboni, I. (2022). Myers-briggs personality classification from social media text using pre-trained language models. *arXiv preprint arXiv:2207.04476*.

Sharp, B. (2015). Towards a cognitive natural language processing perspective. In N. Gala, R. Rapp, & G. BelEnguix (Eds.), *Language production, cognition, and the lexicon* (pp. 25–35, Vol. 48). https://doi.org/10.1007/978-3-319-08043-7_3

Shirazi, F., & Mohammadi, M. (2019). A big data analytics model for customer churn prediction in the retiree segment. *International Journal of Information Management*, *48*, 238–253.

Silva, C. (2003). The importance of stop word removal on recall values in text categorization. *3*, 1661–1666 vol.3. https://doi.org/10.1109/IJCNN.2003.1223656

Srivastava, D., & Bhambhu, L. (2010). Data classification using support vector machine. *Journal of Theoretical and Applied Information Technology*, *12*, 1–7.

Utami, N. A., Maharani, W., & Atastina, I. (2021). Personality classification of facebook users according to big five personality using svm (support vector machine) method [5th International Conference on Computer Science and Computational Intelligence 2020]. *Procedia Computer Science*, *179*, 177–184. https://doi.org/https://doi.org/10.1016/j.procs.2020.12.023

Van Rossum, G. (2020). *The python library reference, release 3.8.2*. Python Software Foundation.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer-Verlag New York Inc.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Verhoeven, B., Daelemans, W., & Plank, B. (2016). Twisty: A multilingual twitter stylometry corpus for gender and personality profiling. *Proceedings of the Tenth international conference on language resources and evaluation (LREC'16)*, 1632–1637.

Wahba, Y., Madhavji, N., & Steinbacher, J. (2022). A comparison of svm against pre-trained language models (plms) for text classification tasks. *International Conference on Machine Learning, Optimization, and Data Science*, 304–313.

Wang, P., Yan, M., Zhan, X., Tian, M., Si, Y., Sun, Y., Jiao, L., & Wu, X. (2021). Predicting self-reported proactive personality classification with weibo text and short answer text. *IEEE ACCESS*, *9*, 77203–77211. https://doi.org/10.1109/ACCESS.2021.3078052

Wu, X., Lin, W., Wang, Z., & Rastorgueva, E. (2020). Author2vec: A framework for generating user embedding. *arXiv preprint arXiv:2003.11627*.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, *135*, 364–381. https://doi.org/https://doi.org/10.1016/j.future.2022.05.014

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al. (2020). Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, *33*, 17283–17297.