UNIVERSITI TEKNOLOGI MARA

COMPARING NLP TEXT PRE-PROCESSING TECHNIQUE WITH CLASSIFICATION OF MBTI PERSONALITY USING MACHINE LEARNING

MUHAMMAD HAZIQ BIN ABDUL RAUF

MSc

February 2023

UNIVERSITI TEKNOLOGI MARA

COMPARING NLP TEXT PRE-PROCESSING TECHNIQUE WITH CLASSIFICATION OF MBTI PERSONALITY USING MACHINE LEARNING

MUHAMMAD HAZIQ BIN ABDUL RAUF

Report submitted in 10/02/2023
of the requirements for the degree of
Master of Computer Science

Faculty of Computer and Mathematical Sciences

February 2023

CONFIRMATION BY PANEL OF EXAMINERS

I certify that a Panel of Examiners has met on 28/01/2023 to conduct the final examination of Muhammad Haziq bin Abdul Rauf in Master in Computer Science. thesis entitled "Comparing NLP Text Pre-processing Technique with Classification of MBTI Personality" in accordance with Universiti Teknologi MARA Act 1976 (Akta 173). The Panel of Examiner recommends that the student be awarded the relevant degree. The Panel of Examiners was as follows:

Ahmad Hamdan Abdullah, PhD
Professor
Faculty of Applied Sciences
Universiti Teknologi MARA
(Chairman)

Mohd Adri Izzat Hamdan, PhD
Associate Professor
Faculty of Applied Sciences
Universiti Teknologi MARA
(Internal Examiner)

Dr Zainura Idrus
Associate Professor
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
(External Examiner)

PM Dr Zalilah Abd Aziz
Professor
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
(Supervisor)

PROF DR HJH HASLINDA YUSOFF
Dean
Institute of Graduates Studies
Universiti Teknologi MARA
Date: 10 January 2020

AUTHOR'S DECLARATION

I declare that the work in this report. was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the results of my own work, unless otherwise indicated or acknowledged as referenced work. This thesis has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that I have been supplied with the Academic Rules and Regulations for Post Graduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Name of Student	:	Muhammad Haziq bin Abdul Rauf
Student I.D. No.	:	2020662592
Programme	:	Master in Computer Science
Faculty	:	Computer and Mathematical Sciences
Report	:	Chemometrics And Pattern Recognition Methods With Applications To Environmental And Biological Studies
Signature of Student	:	 … ……
Date	:	February 2023

ABSTRACT

The goal of this study is to analyze and compare the effectiveness of three NLP strategies for categorizing the Myers-Briggs Type Indicator (MBTI) personality types using written text. Bag of Words (BoW), Word Embedding, and Transfer Learning are the methods that have been assessed for this text classification. TF-IDF is used for BoW while Word2Vec is used for word embedding and lastly for transfer learning its BERT. The method used in this study is the NLP process which is data collection, data cleaning, data processing, feature extraction in which it differs based on type of model between all three NLP strategies, model training and evaluation and lastly is the comparison between the three techniques. The findings from comparing the three different methods of NLP is that transfer learning has the highest accuracy followed by word embedding and lastly BoW. In conclusion, transfer learning has the best accuracy for text classification using MBTI.

ACKNOWLEDGEMENT

Firstly, I wish to thank God for giving me the opportunity to embark on my Master and for completing this long and challenging journey successfully. My gratitude and thanks go to my supervisor PM Dr Zalilah Abd Aziz.

I would like to thank all of the people who helped us with this project, without their support and guidance it wouldn't have been possible. I appreciate PM Dr Zalilah Abd Aziz and Ts Dr Noraini binti Seman for their guidance and supervision which has provided a lot of resources needed in completing our project.

My parents as well as friends were constantly encouraging us throughout the process when we felt discouraged or became frustrated because they knew how much work went into this venture so that is why we want to extend them thanks too!

We are grateful to our colleagues in developing the project, for their willingness and assistance. They helped us with this project, which we appreciate dearly.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviations

NLP	Natural Language Processing
BoW	Bag of Words
BERT	Bidirectional Encoder Representations from Transformers
ML	Machine Learning
SVM	Support Vector Machine
SVC	Support Vector Classifier
RNN	Recurrent Neural Network
TF-IDF	Term Frequency–Inverse Document Frequency
MBTI	Myers Briggs Type Indicator
OCEAN	Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism

Chapter 1

Introduction

1.1 Research background

An individual's personality is a crucial component of their identity since it affects their thoughts, feelings, and actions. A well-known personality evaluation instrument called the Myers-Briggs Type Indicator (MBTI) divides people into one of 16 personality types depending on how they like to express certain traits like introversion vs extroversion, reasoning against feeling, and sense versus intuition. (Myers, I. B., & Briggs, K. C. ,1962). A branch of artificial intelligence and linguistics called "natural language processing" (NLP) is concerned with the use of natural language by computers and people in communication. It is possible to create intelligent systems that can decipher and react to human communication by using NLP techniques to analyze, comprehend, and generate human language. (Li, J., Chen, X., Hovy, E., & Jurafsky, D. ,2015). The MBTI is frequently employed in the areas of team building, career counseling, and personal development. It is a well-liked tool because it offers a structure for comprehending and valuing individual diversity. However, it should be highlighted that because the MBTI lacks solid scientific backing, it should not be utilized as a conclusive indicator of personality. There also has been a lack of research in preprocessing in word vector models. So in this research the researcher will look further into it. (Babanejad, N., Agrawal, A., An, A., & Papagelis, M., 2020, July). The research will compare three text preprocessing techniques which are Bag of Words (BOW), Words Embedding (Word2Vec) and Transfer Learning which uses Bidirectional Encoder Representation from Transformer.

1.2 Motivation

To increase the precision of MBTI prediction models, text preprocessing techniques in NLP for MBTI are being compared. Accurate predictions can have significant applications in

many disciplines, including psychology, education, and human resource management. MBTI is a commonly used personality evaluation tool. Text data must be properly processed and converted in order for it to be relevant for MBTI predictions, including web posts, survey replies, and textual correspondence. To increase prediction accuracy and advance the field of MBTI research utilizing NLP approaches, text preparation techniques must be chosen and compared.

1.3 Statement of problem

The problem is when selecting the best preprocessing technique by comparing three text preprocessing techniques which are BoW, Words Embedding and Transfer Learning to get the best result for classification of MBTI personality.

1.3 Objectives

To find the best text preprocessing technique among Bow, Word Embedding and Transfer Learning in classification of MBTI personality.

To compare the text processing techniques which are BoW, Word Embedding and Transfer Learning.

To make a classification model that is able to determine MBTI personality based on text.

1.3 Significance of study

The research is able to determine the comparison and the best text preprocessing among BoW, Word Embedding and Transfer Learning. The research is also able to make the classification of personality based on MBTI which is widely used in Human Resources to pick a candidate for a job or career counselor to guide students to pick the best career for themselves. (Kin, L. W., & Rameli, M. R. M., 2020)

Chapter 2

Article Review

2.1 Introduction

In this chapter the researcher will do article review based on the research that will be done. There's four section which are personality, machine learning, personality and social media and lastly predicting personality with machine learning.

2.2 Personality

According to Babcock, S. E., & Wilson, C. A. (2020), The five characteristics that make up personality are extraversion, agreeableness, conscientiousness, neuroticism, and openness (OCEAN). How approachable someone is is measured by their extraversion. Being agreeable is being liked by others. How cautious and disciplined a person is typically reflects their conscientiousness. A person's propensity for negative emotions like pessimism, anxiety, and hostility is known as neuroticism. Finally, openness refers to how receptive someone is to novel experiences or concepts. Additionally, it stated that heritable personality, which can account for up to 50% of personality, is a significant determinant.

There are personality theories that focus more on career such as Holland theory and The Myers-Briggs Type Indicator (MBTI). In a systematic review done by Zainudin, Z. N., Rong, L. W., Nor, A. M., Yusop, Y. M., & Othman, W. N. W. (2020) on Holland theory on career decision making, it shows that Holland theory focuses on six types of personality which are Realistic, Investigative, Artistic, Social, Enterprising and Conventional (RIASEC). It considers factors including personality, interests, skills, and surroundings. Only works that apply Holland theory are cited by the author. According to the research, Holland hypothesis is a reliable indicator of profession choice. Other

characteristics have been identified, including financial position, parental support, gender, educational background, and culture.

Woods, R. A., & Hill, P. B. (2020) did a review based on MBTI. MBTI is how a person's different indicators make up their personality and how to perceive as well as how they think.A series of questions is asked based on four indicators which are perceiving, energy, judging and orientation. Based on the four indicators, there are sixteen personality types that can be derived.

The Myers-Briggs Type Indicator (MBTI) is a personality assessment tool that was developed by Isabel Briggs Myers and her mother, Katherine Cook Briggs, based on the psychological theory of Carl Jung. The MBTI categorizes individuals into one of 16 personality types based on their preferences for certain characteristics. These preferences are measured using four dichotomies:

Introversion (I) versus Extroversion (E): This dimension reflects how people prefer to direct their energy and attention. Introverts tend to be more inward-focused and prefer solitude, while extroverts are more outgoing and energized by social interaction.
Sensing (S) versus Intuition (N): This dimension reflects how people prefer to process information. Sensors tend to focus on concrete, observable facts and details, while intuitives tend to focus on abstract, theoretical concepts and patterns.
Thinking (T) versus Feeling (F): This dimension reflects how people prefer to make decisions. Thinkers tend to rely on logic and objective analysis, while feelers tend to consider the emotions and values of themselves and others.
Judging (J) versus Perceiving (P): This dimension reflects how people prefer to structure their lives. Judgers tend to be organized and decisive, while perceivers tend to be flexible and open to new experiences.

Each individual is assigned a four-letter code based on their preferences in these dichotomies. For example, an individual who is introverted, intuitive, thinking, and perceiving would be classified as an INTJ type. (King, S. P., & Mason, B. A., 2020)

Table 1 shows the MBTI personality and the possible personality traits combination

Personality	**Code**	**Possible combination**
Extraversion / Introversion	E/I	ISTJ, ISTP, ISFJ, ISFP, INFJ, INFP, INTJ, INTP, ESTP, ESTJ, ESFP, ESFJ, ENFP, ENFJ, ENTP, ENTJ
Sensing/ Intuition	S/N	
Thinking/ Feeling	T/F	
Judging/ Perceiving	J/P	

Table 1: MBTI personality

2.3 Machine learning

In 'Machine learning and deep learning', Janiesch et al (2021) reported that instead of codifying knowledge into computers, machine learning (ML) seeks to automatically learn. ML aims at automating the task of analytical model building to perform cognitive tasks. Recent advancements in Deep Learning (DL) allow for processing data of different types in combination. This is useful in applications where content is subject to multiple forms of representation. They contribute to the ongoing diffusion of Artificial Intelligence (AI) into electronics markets by discussing four fundamental challenges for intelligent systems based on ML and DL in real-world ecosystems. They contend that they estimate that much of the upcoming research on electronic markets will be against the backdrop of AIaaS and their ecosystems. Related future research will need to address and factor in the challenges they presented. With this fundamentals article, we provide a broad introduction to ML and DL. Often subsumed as AI technology, both fuel the analytical models underlying contemporary and future intelligent systems. We have conceptualized ML, shallow ML, and DL as well as their algorithms and architectures. Further, we have described the general process of automated analytical model building

with its four aspects of data input, feature extraction, model building, and model assessment. Lastly, we contribute to the ongoing diffusion into electronics markets by discussing four fundamental challenges for intelligent systems based on ML and DL in real-world ecosystems.

El Naqa, I., & Murphy, M. J. (2015) did a book on what machine learning is. A key element of digitalization solutions has been machine learning, which is primarily a branch of artificial intelligence. The author will discuss the advantages and disadvantages of machine learning algorithms from the perspective of their applications in order to assist readers in selecting the best learning algorithm. This study attempted a review of the most widely used machine learning techniques to address classification, regression, and clustering issues. The advantages and disadvantages of these algorithms have been discussed, and (when applicable) comparisons of the performance, learning rate, and other characteristics of various algorithms have been made. It has also been suggested how these algorithms might be put to use in actual circumstances. We've discussed semi-supervised learning, unsupervised learning, and supervised learning, three different types of machine learning techniques. It is hoped that by identifying the available machine learning algorithms, readers will be better able to make educated decisions. Machine learning algorithms, then selecting the most suitable algorithm for the current circumstance.

It is possible to create personalized language processing systems that can adjust to the special traits and preferences of individual users by fusing NLP with the MBTI personality theory. This might be used for a variety of things, like creating customized language learning programmes, intelligent tutoring systems, or customer care chatbots that can adjust to the user's communication preferences.

A branch of artificial intelligence and linguistics called "natural language processing" (NLP) is concerned with the use of natural language by computers and people in communication. To enable efficient communication and information processing, NLP aims to give computers the ability to comprehend, interpret, and produce human language. (Cambria, E., & White, B. , 2014)

Language translation, sentiment analysis, text classification, dialogue systems, and information extraction are just a few of the many uses for NLP approaches. Typical NLP assignments include:

Tokenization: the process of breaking down a piece of text into smaller units called tokens, such as words or punctuation marks.

Part-of-speech tagging: the process of labeling each token with its corresponding part of speech (e.g. noun, verb, adjective).

Named entity recognition: the process of identifying and classifying named entities (e.g. people, organizations, locations) in a piece of text.

Stemming: the process of reducing a word to its base form (e.g. jumping -> jump).

Lemmatization: the process of reducing a word to its base form while taking into account the part of speech (e.g. jumping -> jump, jumps -> jump).

This research focuses on the last step of data processing in NLP which is to vectorize the data so that the machine learning model can learn from it. The techniques to vectorize the data in this research are Bag of Words (BoW), Word Embedding and Transfer learning. For BoW, the Term frequency-inverse document frequency (TF-IDF) model is chosen. Whereas for Word Embedding, Word2Vec is the model. Lastly for Transfer Learning, Bidirectional Encoders Representation from Transformer is the model chosen.

NLP has numerous applications in various industries, including healthcare, finance, marketing, and customer service. It has the potential to improve efficiency, accuracy, and productivity by automating language-based tasks and enabling more effective communication between humans and machines. (Mathews, S. M. ,2019)

2.4 Personality and social media

Correa, T., Hinsley, A. W., & De Zuniga, H. G. (2010) report that three-quarters of American adults have been online, with even more teens (93%) reporting they do so. More than half of America's teens and young adults send instant messages and use social networking sites.Extraverted individuals are typically more emotionally stable and receptive to new experiences. The human experience has been significantly altered by the Internet. The literature on the applications of recently adopted technologies in society is advanced by this essay. It investigates the connection between personality traits and user-generated software. The study had 8568 participants in all. Some of their findings may complement what was already known about this topic: The first research question looked at whether there were any gender differences in the relationship between personality predictors and social media use. The findings showed that there were several differences. This agrees with earlier recommendations from other studies. They contend that by distinguishing between related forms of interaction, a broad range of different uses within the social media space would significantly advance this line of research. In the final subsample, there were more female respondents to the survey than male respondents. Extraverted individuals typically exhibit greater emotional stability and are more receptive to novel experiences. There is a positive correlation between life satisfaction and extraversion, $r = .14$, $p < .001$. These results suggest that people who are more extraverted and stable tend to be more content with life. There is no relationship between personal contentment with life and being open to experiences, $r = .02$, n.s

Gwendolyn Seidman (2012) reported in 'Self-presentation and belonging on Facebook' that according to Nadkarni and Hofmann's (2012) dual-factor model, Facebook use is motivated by two needs: belonging and self-presentation. The Big Five traits are openness, conscientiousness, agreeableness, extraversion, and neuroticism. Extraversion is related to several belongingness-related constructs. Agreeable individuals use Facebook to present actual self traits and refrain from attention-seeking. Neuroticism is associated with several outcomes relating to belongingness needs. The Internet has had a major impact on social life. High agreeableness and neuroticism were the best predictors of belongingness. Conscientious individuals are cautious in their online self-presentations. The analysis involved 184 undergraduates. Their results appear to provide a counterpoint to earlier work in this area: "Introverted individuals are more

likely to report using Facebook to keep up with friends. Gosling et al found that extraversion was positively associated with viewing others' Facebook pages. It is unclear how extraversion is related to Facebook to learn about others," Seidman claimed. Discussing potential improvements, A major limitation of this work is reliance on self-report. However, many of the variables assessed in the present study were subjective. The dependent measures were created for the purpose of this study and thus their reliability and validity are not well-established that they concede. The group advocates that future research should examine influences on acceptance-seeking behavior in neurotic individuals.

2.5 Predicting personality with machine learning

Sakdipat Ontoum and Jonathan H. Chan (2022) did research which uses Naive Bayes, RNN and SVM for their machine learning algorithms to predict personality based on MBTI while using CRISPDM (Cross-Industry Standard Process for Data Mining) to guide the learning process which is also an agile methodology. CRISPDM. The first step of CRISPDM is business understanding followed by data understanding, data preparation, modeling, evaluation and lastly deployment. For the data preprocessing, they increase the column by four on MBTI to increase accuracy. Then the words are selected and removed based on what they are. Words that include links are removed while unique words are selected. Then they proceed to do lemmatization which changes words into their root form so that words that have similar meaning are categorized the same. Tokenization is made to give words that are unique more meaning while common words less meaning. Then the data is split into training and testing to train the machine learning algorithm. The result shows that RNN has better accuracy than SVM and Naive Bayes with 49.75% accuracy which is 8% more accuracy than others.

A research group led by Simone Leonardi at the Department of Control and Computer Engineering (DAUIN) (2020) described multilingual transformer-based personality traits estimation. Language models have been widely employed to measure

personality traits starting from written text. They preprocesses their text using CLS token after removing punctuations. They processed this array of 768 features with a stacked neural network to perform a regression on each of the five personality traits. They then pre trained their data using BERT. The team worked in a supervised environment, where expected personality traits are numbers in the continuous range 1–5. They argue that current techniques do not consider polysemy in text and differences among languages. Leonardi and colleagues want to compute them in real time during the dialogue to improve the digital experience. The group worked on a subset of the whole dataset made of 9913 samples defined as myPersonality small. They used two files in the dataset: in the first one, each line is made up of a message id, a user id, the plain text of message, and other information. The group argues that they describe a model to process social media posts, encoding each of them on a sentence level into a high-dimensional array space. The team processed this array of 768 features with a stacked neural network to perform a regression on each of the five personality traits in the Big 5 model. They also improved the model by doing hyperparameters and weight tuning on each of personality traits separately. Overall there is an improvement in the result of about 30% of each of the personality traits.

Alam Sher Khan et al (2020) did a paper on personality classification from online text using a machine learning approach. Text from social media is used to predict and describe an individual's behaviour and influences daily life activities. Increasing use of Social Networking Sites, such as Twitter and Facebook have propelled the online community to share ideas, sentiments, opinions, and emotions with each other, reflecting their attitude, behavior and personality. They used random over-sampling to balance their dataset. They use tokenization, dropping stop words, word stemming, feature selection via TF-IDF and count vectorizer to preprocess their text data. The overall comparison of predicting personality traits is presented using all evaluation metrics to determine the performance of different classifiers. The results show that XGBOOST in collaboration with LIWC and TF-IDF feature vectors gave accurate prediction scores for all four traits. There were 50 social media posts included in the research. Aspects of the authors' results look to diverge from earlier research in the field: "SVM in collaboration with LIWC and TF-IDF feature vectors gave accurate prediction scores for all four traits. MLP with all features

Vectors got maximum accuracy score for S/N trait however its result for J/P trait is lower," Khan claimed. Khan got 99% precision and accuracy forI/E and Sensing vs Intuition traits and obtained about 95% accuracy for Thinking vs Feeling and Judging vs Perceiving dimensions using XGBOOST. There are also other algorithms used such as KNN, Decision Tree, Random Forest, Logistic Regression, MLP classifier and SVM.

Mohammad Hossein Amirhosseini and Hassan Kazemian (2019) focus their study on using XGBOOST to predict personality traits based on MBTI. The text processing is done using NLTK which is a natural language processing tools library in Python. NLTK does the classification, tokenization, stemming, tagging, parsing and semantic reasoning. The researcher also processed characters and symbols as they considered it a part of the text. The special characters are processed using Part-Of-Speech tagging. They compared their findings based on previous research that uses other machine learning algorithms such as SVM, Naive Bayes and KNN. They argued Extreme Gradient Boosting shows better performance due to the use of more regularized model formalization in order to control over-fitting. The algorithm is based on combining rough inaccurate rules of thumb with principles with broad application that are not intended to be strictly accurate or reliable. They did several experiments on XGBOOST based on hyperparameter tuning such as tweaking max depths, nthreads, n_estimators and learning_rate. As well as parameters such as number and size of trees, the number of trees and learning rate as well as the row and column subsampling rates. The accuracy of XGBOOST was compared with RNN and XGBOOST showed a better performance.

A paper done by Nur Haziqah Zainal Abidin et al (2020) with the goal of building a prediction system that can automatically predict the users' personality based on their activities on Facebook and twitter. In this research, the researcher will only use Random Forest classifiers for prediction of personality. It is then compared with the previous result on the same dataset with different machine learning algorithms such as Logistic regression, KNN and SVM. The data is first pre-processed by removing special characters. The data is then pre-processed using TF-IDF and Word2Vec which is a preprocessing step of changing textual data into numeric data. The result shows that Random forest algorithms have better

results compared to other algorithms based on Pearson Correlation with the correlation between all Extraversion vs Introversion, Sensing vs Intuition, Judging vs Perception and Feelers vs Thinkers which are at 100%. The closest is KNN followed by logistic regression and the worst is SVM. The result is better with Word2Vec compared to TF-IDF.

Dharshni, P. et al (2021) did research on personality prediction based on user behaviour on social media. This research uses MBTI as their psychological feature. The data is collected from Web application design course student's micro blogging tools which they used for collaboration activities and communication in a project based learning scenario. The three machine learning algorithms used in this research are logistic regression, SGD classifier and KNN. SGD classifier is a convex machine learning algorithm in its loss function, this reduces the cost function. It has smoothness properties in its iterative function. The result shows that it has high accuracy for its I/E and S/N traits which is 99% and 95% accuracy for both T/F and J/P traits.

Gokalp Mavis, Ismail Hakki Toroslu and Pinar Karagoz (2021) did a study on personality prediction using classification on Turkish tweets. The personality theory used is the big five personality traits (OCEAN). The study used both supervised machine learning and deep learning. The supervised machine learning algorithm used is Random forest, SVC, linear SVC and KNN. Before running the data through supervised learning, the data is preprocessed by filtering out tweets that have little to no meaning. Special characters are removed. Then the pre-processing is done using Term frequency - Inverse Document Frequency (TF-IDF) and Word2Vec to get its vector so that supervised machine learning can process it and run it through its algorithm. The deep learning model used is Long short term memory (LTSM) which is an artificial neural network. The data is collected through surveys and the response is calculated from the value of 0 to 100 for each personality trait. Other features such as use of language, timestamps, emoticons and non-text twitter user information is also used. Then the data went through feature elimination and normalization to get more optimized data. For supervised machine learning, the researcher used TF-IDF weighting to remove unnecessary and redundant data. Then the data is changed in vector form with Word2Vec. The experiment is run with the selected features only from the

turkish dataset which is only 20% of the english dataset. Another experiment is run with all the features. The last experiment was run with LSTM. For supervised machine learning, SVC shows consistency as the highest accuracy in supervised machine learning model. For both machine learning model and deep learning model the highest accuracy is for agreeableness which is the highest at .88 for deep learning. The lowest accuracy for conscientiousness at .44 for reduced features supervised machine learning. LSTM shows higher accuracy for neuroticism and agreeableness but lower in openness, conscientiousness and extraversion.

A study done by Keh, S. S., & Cheng, I. (2019) which uses pre-trained models to make an MBTI personality classification. The research uses data from personalitycafe.com. The researcher also tries to make a personality classification system. The pre-trained model used for the research is BERT. For data pre processing, the special characters and punctuations. After that the text was changed to all lowercase letters. Then tokenization is done using BERT tokenizer. The text was trained using BERT and adjustment of parameters is done. After the adjustment of parameters such as learning rate and number of epochs, the accuracy increased from .09 to .48.

Lastly, a study on comparative analysis of machine learning models in personality predictions by Nisha, K. A., Kulsum, U., Rahman, S., Hossain, M. F., Chakraborty, P., & Choudhury, T. (2022). In this research, the personality used were the Big five personality. The text processing done are URL , extra character removals , punctuation remover, tokenization, stop words removal, stemming and lastly CountVectorizer. The classification models used for this research are Naive Bayes, SVM and XGBoost. SVM and XGBoost performed quite well while Naive Bayes did worse than the two. The accuracy is said to be improved when using a balanced dataset.

Chapter 3
Methodology

3.1 Introduction

In this chapter, the researcher will present on how the study will be conducted. The researcher will construct a methodology on how to achieve the goal of this study. This chapter is divided into several part such as research approach, research design, the population and the sample, research instrument, data collection, data analysis, validity and reliability.

3.2 Research approach

The research approach for this study will be a descriptive qualitative approach because the researcher is trying to classify and predict the personality of a person based on what they post on personalitycafe.com . Descriptive research is a research done, where the researcher cannot control the variables. It can be distinguished by trying to explain, determine or identify the relationships of the variable (Ethridge, 2004). This is suitable for this research since the researcher wants to do machine learning which involves data mining. The researcher will try to predict what is the person's personality in MBTI based on what they posted on personalitycafe.com. The variables will be personality and the personality types based on MBTI 16 personality types.

3.3 Research design

The design of this research is in quantitative mode since the data collection is done through the posts on personalitycafe.com via google bigquery. Quantitative mode in this research is a method that focuses on statistical analysis of the data that have been collected using google bigquery run through machine learning to predict the new data collected. Quantitative research focuses on collecting data and deducing it towards the groups of people (Babbie and earl, 2010).

The research will be done by having personality and text posted as the variable. The user that posted on personalitycafe.com is the target for sample collection.

3.4 Population and sample

This research objective is done to predict the personality of a person based on the text they posted on social media. So the sample must be the population of people who posted their text on social media. Convenience random sampling will be done on the person who posted on personalitycafe.com. Random sampling is a selection technique where the sample is collected randomly but the selection probability is known (Trobia .A, Lavrakas, 2008). The strength of is that it is easy to do, cheapest and least time taken but the weakness is that there might be selection bias (Taherdoost, 2016). The sample will be collected by chance. It is collected from a population that is available for the researcher.

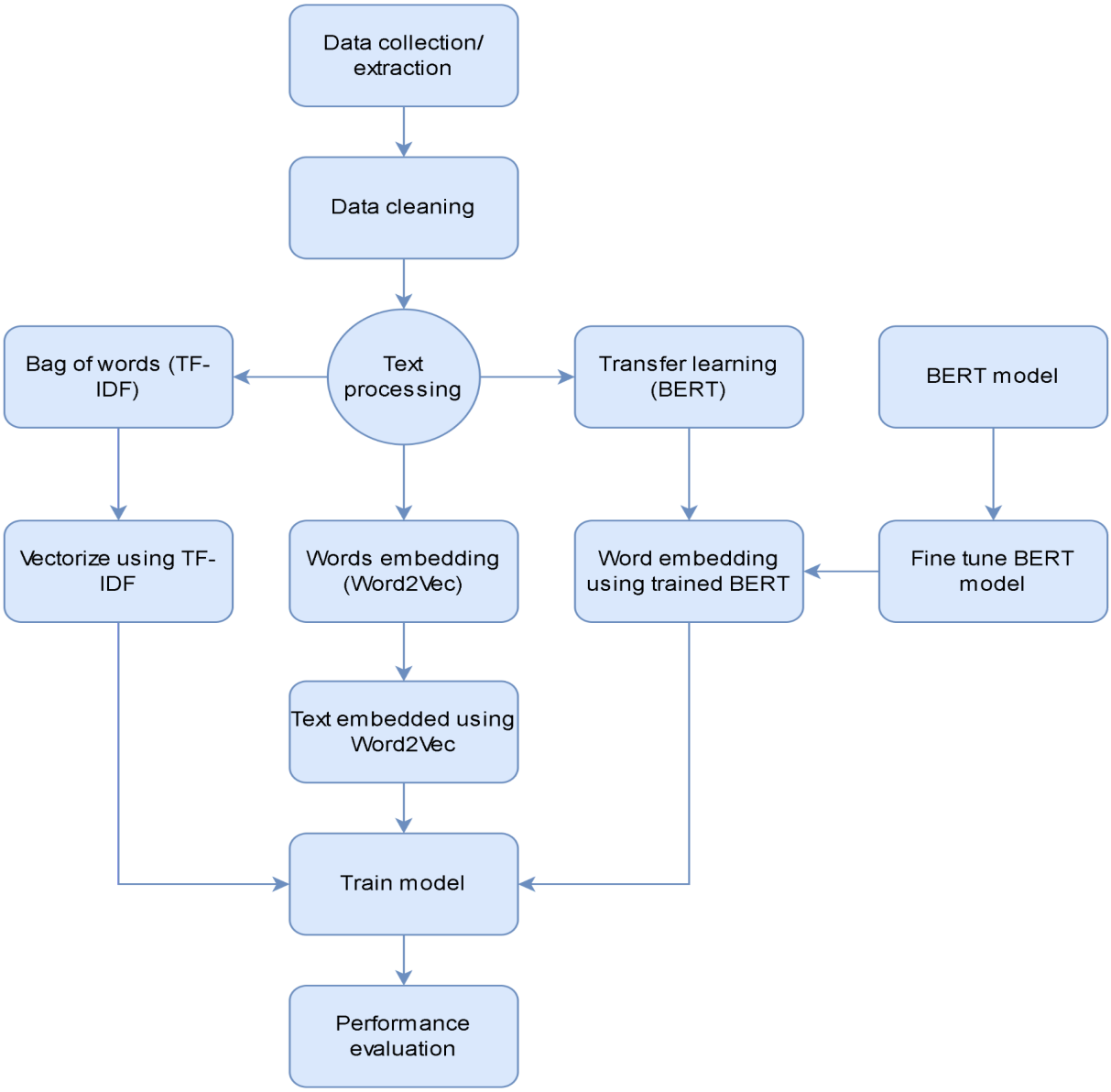

Figure 1: Flowchart of methodology

3.5 Data collection

Data is collected using google bigquery. The data consisted of four columns which are index, author flair (username / personality), body or text, subreddit. The data is collected from various personality types based on sixteen MBTI personality types.

3.6 Data cleaning

For NLP there are multiple steps for data preprocessing. The first step is cleaning the data. The researcher would remove outlier data such as missing, duplicate and anomalies. Columns that are not needed such as index and subreddit are removed as well. The researcher then balanced the dataset due to the data being too unbalanced with the most amount of a single feature reaching more than 400 000 and the lowest is less than 25 000. The data is balanced using an undersampling method which reduces all the number of features to match the lowest number of features.

3.7 Data pre-processing

Based on research of Smelyakov, K., Karachevtsev, D., Kulemza, D., Samoilenko, Y., Patlan, O., & Chupryna, A. (2020, October) data processing follows a flow of turning all the text to lowercase letters and remove punctuation as well as characters. This step is done using regular expressions. Next step is the removal of stopwords. Stopwords are the words that hold low-level information and make other texts with more important information more focused on.

The next step is tokenization which is breaking a paragraph into smaller parts, sentences or strings. These smaller parts can be further tokenized into smaller parts such as words. After that is lemmatization which is converting words into root words such as caring into care. This step is further improved with stemming which is the removal of affixes to the root word. Affixes such as -ing in flying and -ly in readily. Stemming is done in the absence of words that are not lemmatized. Lemmatize gives meaning to the words that are processed. While stemming, the

word might lose its meaning. Lemmentize is more computationally expensive than stemming. (Saumyab271,28 June 2022)

3.8 Data pre-processing technique

The first data pre-processing technique is BOW which is counting the number in which the word occurs and putting it into a vector. This will cause the word to lose its meaning. (Juluru, K., Shih, H. H., Keshava Murthy, K. N., & Elnajjar, P., 2021). The technique used for this BOW is TF-IDF. TF-IDF is is a metric used in statistics to determine how pertinent a word is to a particular document among a group of documents. A word's frequency in a document and its inverse document frequency across a set of documents are multiplied in order to achieve this. (Kim, S. W., & Gil, J. M., 2019)

Other than that, the use of word embedding in this research is Word2Vec. Word embedding is putting meaning into each word and vectorizing it. For Word2Vec, it's based on a one-hot encoded label of input to predict the output. Word2Vec is the use of either skip-gram and Common Bag of Words (CBOW) and both is a single neural network. In this research the skip-gram method is used. (Jang, B., Kim, I., & Kim, J. W., 2019)

Lastly, the technique used in the research is Transfer Learning specifically using BERT verison DISTILLBERT. Based on research done by Alzahrani, E., & Jololian, L. (2021), BERT required the data to be pre-trained using BERT, the BERT model will use the data that is cleaned and preprocessed with its own model. After the data is pre-trained with BERT then the researcher can continue to use the pre-trained model to transfer it to their own model hence the name transfer learning.

3.9 Machine learning model

The machine learning model used for this research is logistic regression. Logistic regression is a predictive analytics and classification that frequently makes use of this kind of statistical model, also referred to as a logit model. Based on a given dataset of independent

variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. Given that the result is a probability, the dependent variable's range is 0 to 1. In logistic regression, the odds—that is, the probability of success divided by the probability of failure—are transformed using the logit formula. The following formulas are used to represent this logistic function, which is also referred to as the log odds or the natural logarithm of odds:

$$\text{Logit}(pi) = 1/(1+ \exp(-pi))$$

$$\ln(pi/(1-pi)) = \text{Beta}\_0 + \text{Beta}\_1*X\_1 + \ldots + B\_k*K\_k \qquad (1)$$

Logit(pi) is the dependent or response variable in this logistic regression equation while x is the independent variable. The most common method for estimating the beta parameter, or coefficient, in this model is maximum likelihood estimation (MLE). In order to find the best fit for the log odds, this method iteratively tests various beta values. The log likelihood function is created after each of these iterations, and logistic regression aims to maximise this function to find the most accurate parameter estimate. The conditional probabilities for each observation can be calculated, logged, and added together to produce a predicted probability once the best coefficient (or coefficients, if there are multiple independent variables) has been identified. If the classification is binary, a probability of less than.5 predicts 0 and a probability of more than 0 predicts 1. It is recommended to assess the model's goodness of fit, or how well it predicts the dependent variable, after the model has been computed. The Hosmer-Lemeshow test is a well-liked technique for evaluating model fit.

3.10 Result of classification

The result is analyzed based on recall, precision and accuracy. The classification report that is plotted to see these results.

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}} \qquad (2)$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}} \qquad (3)$$

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}} \qquad (4)$$

As the formula above we can say that, precision is how precise/accurate our model is will depend on how many of the predicted positive results are actually positive. Precision is focused only on one feature, the same as recall. For recall, it determines how many Actual Positives our model actually captures by classifying it as Positive. Lastly, accuracy is the accurate prediction of the model based on the actual number of data.

Chapter 4

Result and discussion

4.1 Introduction

In this chapter, the researcher will talk about the result that is achieved from the research based on the methodology. The result of BOW, Word Embedding and Transfer Learning is analyzed based on their accuracy, recall and precision.

4.2 BOW (TF-IDF) result

The first model that was trained is the model that uses BOW as its text processing. There is an issue where the recall is high only for ENFP and for the others its 0 except for ISTP. The precision only have values for ENFP and ISTP as well. Meaning that it's very good to capture the personality of ENFP and bad at ISTP while not able to capture anything for the others. The accuracy is also low at .05.

Accuracy: 0.05
Auc: 0.5
Detail:

	precision	recall	f1-score	support
ENFJ	0.00	0.00	0.00	67
ENFP	0.05	0.97	0.10	61
ENTJ	0.00	0.00	0.00	69
ENTP	0.00	0.00	0.00	79
ESFJ	0.00	0.00	0.00	81
ESFP	0.00	0.00	0.00	70
ESTJ	0.00	0.00	0.00	74
ESTP	0.00	0.00	0.00	70
INFJ	0.00	0.00	0.00	73
INFP	0.00	0.00	0.00	83
INTJ	0.00	0.00	0.00	78
INTP	0.00	0.00	0.00	62
ISFJ	0.00	0.00	0.00	93
ISFP	0.00	0.00	0.00	79
ISTJ	0.00	0.00	0.00	77
ISTP	0.18	0.04	0.06	84
accuracy			0.05	1200
macro avg	0.01	0.06	0.01	1200
weighted avg	0.01	0.05	0.01	1200

Figure 3: Classification report for BOW

4.2 Word Embedding (Word2Vec) result

For Word Embedding, we can see improvement in accuracy, recall and precision. In which accuracy has improved from .05 to .078. There is also an improvement in recall which there value in almost all the feature output as well as for precision.

Accuracy: 0.078

	precision	recall	f1-score	support
ENFJ	0.12	0.12	0.12	64
ENFP	0.11	0.04	0.06	67
ENTJ	0.06	0.11	0.08	56
ENTP	0.07	0.16	0.10	56
ESFJ	0.14	0.08	0.10	65
ESFP	0.08	0.10	0.09	63
ESTJ	0.11	0.15	0.13	54
ESTP	0.09	0.09	0.09	65
INFJ	0.16	0.05	0.07	64
INFP	0.00	0.00	0.00	79
INTJ	0.00	0.00	0.00	75
INTP	0.11	0.04	0.06	70
ISFJ	0.06	0.02	0.02	64
ISFP	0.05	0.15	0.08	53
ISTJ	0.05	0.20	0.08	45
ISTP	0.07	0.05	0.06	60
accuracy			0.08	1000
macro avg	0.08	0.08	0.07	1000
weighted avg	0.08	0.08	0.07	1000

Figure 4: Classification report for Word Embedding

4.3 Transfer Learning (BERT)

For transfer learning the result was the best among the three. There's value in almost all recall and precision except for INTJ. The accuracy is the highest among the others with .1. The recall is balanced except for ESTP which is higher and INTJ which is 0.

Accuracy: 0.10833333333333334

	precision	recall	f1-score	support
ENFJ	0.07	0.04	0.06	45
ENFP	0.11	0.09	0.10	46
ENTJ	0.09	0.18	0.12	22
ENTP	0.03	0.03	0.03	36
ESFJ	0.08	0.11	0.09	37
ESFP	0.09	0.11	0.10	37
ESTJ	0.27	0.16	0.20	44
ESTP	0.23	0.24	0.24	33
INFJ	0.10	0.14	0.12	35
INFP	0.07	0.05	0.06	37
INTJ	0.00	0.00	0.00	31
INTP	0.09	0.11	0.10	37
ISFJ	0.06	0.05	0.06	38
ISFP	0.10	0.07	0.08	43
ISTJ	0.24	0.14	0.18	43
ISTP	0.21	0.25	0.23	36
accuracy			0.11	600
macro avg	0.11	0.11	0.11	600
weighted avg	0.12	0.11	0.11	600

Figure 5: Classification report for Transfer Learning

4.4 Discussion

The objective for this research is achieved where we can see that there's a significant difference between all the text processing techniques. The highest which is Transfer Learning has double the accuracy compared to BOW and the recall and precision is more evenly distributed meaning that it can classify the result for most of the features. There may need improvements to make the result better such as increasing the amount of data used or trying other machine learning algorithms. Machine learning algorithms such as neural networks and random forests may get better results.

Chapter 5

Conclusion

5.1 Introduction

In this chapter the researcher will give the conclusion for this research. From introduction to result and discussion. For chapter 1, the researcher the research background, problem statement, objectives and significance of the study. Other than that for chapter 2 the researcher did article review on MBTI personality, machine learning, personality and social media, predicting / classifying personality with machine learning and text processing technique. Then for chapter 3, the researcher talked about the methodology of this research. Started with data collection, then data cleaning, text preprocessing, text preprocessing technique and lastly machine learning model used. Lastly in chapter 4, the researcher analyzes the data that is gained from the research based on text preprocessing technique.

5.2 Objective Outcome

The first objective of this study is achieved which is comparing the three text preprocessing technique and finding the best within the three text preprocessing technique which were BoW, Word Embedding and Transfer Learning. Transfer learning yield the best accuracy.

For the second objective, the classification model is able to classify the MBTI personality though the accuracy, recall and precision is low.

5.2 Recommendations

For recommendations, the researcher suggests the use of other machine learning models such as random forest since random forest or bayesian network might be better for classification. Other than that, is the use of deep learning models. Other than that, the researcher

recommends the use of more data for the model to learn more. Furthermore, the researcher suggest to focus on either semantic analysis or sentiment analysis. Lastly, the researcher suggest the tuning of hyperparamenter such as learning rate, type of optimizer used.

REFERENCES

Amirhosseini, M. H., & Kazemian, H. (2019). Automating the process of identifying the preferred representational system in Neuro Linguistic Programming using Natural Language Processing. *Cognitive processing*, *20*(2), 175-193.

Abidin, N. H. Z., Remli, M. A., Ali, N. M., Phon, D. N. E., Yusoff, N., Adli, H. K., & Busalim, A. H. (2020). Improving intelligent personality prediction using Myers-Briggs type indicator and random forest classifier. *International Journal of Advanced Computer Science and Applications*, *11*(11).

Alzahrani, E., & Jololian, L. (2021). How different text-preprocessing techniques using the bert model affect the gender profiling of authors. *arXiv preprint arXiv:2109.13890*.

Babbie, E. (2010). Research design. The practice of social research, 12

Babanejad, N., Agrawal, A., An, A., & Papagelis, M. (2020, July). A comprehensive analysis of preprocessing for word representation learning in affective tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5799-5810).

Babcock, S. E., & Wilson, C. A. (2020). Big Five Model of Personality. The Wiley Encyclopedia of Personality and Individual Differences: Personality Processes and Individual Differences, 55-60.

Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, *9*(2), 48-57.

Connelly, L. (2020). Logistic regression. *Medsurg Nursing*, *29*(5), 353-354.

Correa, T., Hinsley, A. W., & De Zuniga, H. G. (2010). Who interacts on the Web?: The intersection of users' personality and social media use. *Computers in human behavior*, *26*(2), 247-253.

Dharshni, P., Pon, J., Monisha, M., Nayanthara, C., & Krishna Priya, G. Personality Prediction Based on User Behavior on Social Media.

El Naqa, I., & Murphy, M. J. (2015). *What is machine learning?* (pp. 3-11). Springer International Publishing.

Ethridge, D. (2004). *Research methodology in applied economics: organizing, planning, and conducting economic research* (No. BOOK). Blackwell publishing.

Jang, B., Kim, I., & Kim, J. W. (2019). Word2vec convolutional neural networks for classification of news articles and tweets. *PloS one*, *14*(8), e0220976.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685-695.

Juluru, K., Shih, H. H., Keshava Murthy, K. N., & Elnajjar, P. (2021). Bag-of-words technique in natural language processing: a primer for radiologists. *Radiographics*, *41*(5), 1420-1426.

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

Khan, A. S., Hussain, A., Asghar, M. Z., Saddozai, F. K., Arif, A., & Khalid, H. A. (2020). Personality classification from online text using machine learning approach. *International journal of advanced computer science and applications*, *11*(3).

Kim, S. W., & Gil, J. M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. *Human-centric Computing and Information Sciences*, *9*, 1-21.

King, S. P., & Mason, B. A. (2020). Myers‐Briggs type indicator. The Wiley Encyclopedia of Personality and Individual Differences: Measurement and Assessment, 315-319.

Kin, L. W., & Rameli, M. R. M. (2020). Myers-Briggs Type Indicator (Mbti) personality and career indecision among malaysian undergraduate students of different academic majors. Universal Journal of Educational Research, 8(5A), 40-45.

Leonardi, S., Monti, D., Rizzo, G., & Morisio, M. (2020). Multilingual transformer-based personality traits estimation. *Information*, *11*(4), 179.

Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2015). Visualizing and understanding neural models in nlp. arXiv preprint arXiv:1506.01066.

Mathews, S. M. (2019). Explainable artificial intelligence applications in NLP, biomedical, and malware classification: a literature review. In *Intelligent Computing: Proceedings of the 2019 Computing Conference, Volume 2* (pp. 1269-1292). Springer International Publishing.

Mavis, G., Toroslu, I. H., & Karagoz, P. (2021). Personality analysis using classification on Turkish tweets. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, *15*(4), 1-18.

Myers, I. B. (1962). The Myers-Briggs Type Indicator: Manual (1962).

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. In *Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2021* (pp. 13-23). Springer Singapore.

Ontoum, S., & Chan, J. H. (2022). Personality Type Based on Myers-Briggs Type Indicator with Text Posting Style by using Traditional and Deep Learning. *arXiv preprint arXiv:2201.08717*.

Saumyab271 (,28 June 2022), Stemming vs Lemmatization in NLP: Must-know difference, https://www.analyticsvidhya.com/blog/2022/06/stemming-vs-lemmatization-in-nlp-must-know-differences/#:~:text=Lemmatization%20considers%20the%20context%20and,where%20performance%20is%20an%20issue.

Seidman, G. (2013). Self-presentation and belonging on Facebook: How personality influences social media use and motivations. *Personality and individual differences*, *54*(3), 402-407.

Smelyakov, K., Karachevtsev, D., Kulemza, D., Samoilenko, Y., Patlan, O., & Chupryna, A. (2020, October). Effectiveness of preprocessing algorithms for natural language processing applications. In *2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T)* (pp. 187-191). IEEE.

Taherdoost, H. (2016). Sampling methods in research methodology; how to choose a sampling technique for research. *How to choose a sampling technique for research (April 10, 2016)*.

Trobia, A., & Lavrakas, P. (2008). Encyclopedia of survey research methods.

Woods, R. A., & Hill, P. B. (2020). Myers Brigg. StatPearls [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK554596/#article-25455.s6

Zainudin, Z. N., Rong, L. W., Nor, A. M., Yusop, Y. M., & Othman, W. N. W. (2020). The Relationship of Holland Theory in Career Decision Making: A Systematic Review of Literature. Journal of Critical Reviews, 7(9), 884-892.

Appendices

```
'''
Preprocess a string.
:parameter
    :param text: string - name of column containing text
    :param lst_stopwords: list - list of stopwords to remove
    :param flg_stemm: bool - whether stemming is to be applied
    :param flg_lemm: bool - whether lemmitisation is to be applied
:return
    cleaned text
'''
def utils_preprocess_text(text, flg_stemm=False, flg_lemm=True, lst_stopwords=None):
    ## clean (convert to lowercase and remove punctuations and characters and then strip)
    text = re.sub(r'[^\w\s]', '', str(text).lower().strip())

    ## Tokenize (convert from string to list)
    lst_text = text.split()
    ## remove Stopwords
    if lst_stopwords is not None:
        lst_text = [word for word in lst_text if word not in
                    lst_stopwords]

    ## Stemming (remove -ing, -ly, ...)
    if flg_stemm == True:
        ps = nltk.stem.porter.PorterStemmer()
        lst_text = [ps.stem(word) for word in lst_text]

    ## Lemmatisation (convert the word into root word)
    if flg_lemm == True:
        lem = nltk.stem.wordnet.WordNetLemmatizer()
        lst_text = [lem.lemmatize(word) for word in lst_text]

    ## back to string from list
    text = " ".join(lst_text)
    return text
```

Code for creating function for Tokenization, Removal for stopwords, Stemming and lemmatization

```
df["text_clean"] = df["text"].apply(lambda x:
          utils_preprocess_text(x, flg_stemm=False, flg_lemm=True,
          lst_stopwords=lst_stopwords))
df.head()
```

Code for running function to preprocess text

```
## Tf-Idf (advanced variant of BoW)
vectorizer = feature_extraction.text.TfidfVectorizer(max_features=10000, ngram_range=(1,2))
```

Code for running vectorizer (TF-IDF)

```
w2v_model = gensim_api.load("word2vec-google-news-300")

[==================================================] 100.0% 1662.8/1662.8MB downloaded


def embedding_feats(list_of_lists):
    DIMENSION = 300
    zero_vector = np.zeros(DIMENSION)
    feats = []
    for tokens in list_of_lists:
        feat_for_this =  np.zeros(DIMENSION)
        count_for_this = 0 + 1e-5 # to avoid divide-by-zero
        for token in tokens:
            if token in w2v_model:
                feat_for_this += w2v_model[token]
                count_for_this +=1
        if(count_for_this!=0):
            feats.append(feat_for_this/count_for_this)
        else:
            feats.append(zero_vector)
    return feats

train_vectors = embedding_feats(texts_processed)
```

Code to run Word Embedding (Word2Vec)

```
# For DistilBERT:
model_class, tokenizer_class, pretrained_weights = (ppb.DistilBertModel, ppb.DistilBertTokenizer, 'distilbert-base-uncased')

# Load pretrained model/tokenizer
tokenizer = tokenizer_class.from_pretrained(pretrained_weights)
model = model_class.from_pretrained(pretrained_weights)
```

Code for variable for BERT

```
tokenized = batch_1["text_clean"].apply((lambda x: tokenizer.encode(x, add_special_tokens=True)))


max_len = 0
for i in tokenized.values:
    if len(i) > max_len:
        max_len = len(i)

padded = np.array([i + [0]*(max_len-len(i)) for i in tokenized.values])


np.array(padded).shape

(2400, 186)


attention_mask = np.where(padded != 0, 1, 0)
attention_mask.shape

(2400, 186)


input_ids = torch.tensor(padded)
attention_mask = torch.tensor(attention_mask)
```

Code for BERT tokenization, and to make text into tensor so it can be pre-trained with BERT.

```
import tensorflow as tf
print("Tensorflow version " + tf.__version__)

try:
  tpu = tf.distribute.cluster_resolver.TPUClusterResolver()  # TP U detection
  print('Running on TPU ', tpu.cluster_spec().as_dict()['worker'])
except ValueError:
  raise BaseException('ERROR: Not connected to a TPU runtime; please see the previous cell in this notebook for instructions!')

tf.config.experimental_connect_to_cluster(tpu)
tf.tpu.experimental.initialize_tpu_system(tpu)
tpu_strategy = tf.distribute.TPUStrategy(tpu)
use_gpu = True
# Create model
with tpu_strategy.scope():
    model = model
if not use_gpu:
    model = model

with torch.no_grad():
    last_hidden_states = model(input_ids, attention_mask=attention_mask)
```

Code to run BERT

```
classifier = LogisticRegression()
train_data, test_data, train_cats, test_cats = train_test_split(train_vectors, df["y"])
classifier.fit(train_data, train_cats)
print("Accuracy: ", classifier.score(test_data, test_cats))
preds = classifier.predict(test_data)
print(classification_report(test_cats, preds))
```

Code for Logistic regression and classification report