DETERMINING USER TYPES FROM TWITTER ACCOUNT CONTENT AND STRUCTURE

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

MESUT GÜRLEK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

MARCH 2021

ABSTRACT

DETERMINING USER TYPES FROM TWITTER ACCOUNT CONTENT AND STRUCTURE

Gürlek, Mesut
M.S., Department of Computer Engineering
Supervisor: Prof. Dr. Ismail Hakkı Toroslu

March 2021, 58 pages

People are using social media platforms more and more every day; hence, they are becoming suitable for research studies by their rich content. Twitter is one of the biggest and most widely used social media platforms, and many studies focus on Twitter for social media research. In this thesis, we propose methodologies for determining user types of Twitter accounts by their metadata, content, and structure. Our first problem is classifying organization vs. individual account types using only metadata. After we give details about the data collection and analysis process, we clean text features and normalize number features. Proposed model contains LSTM for textual features and MLP for numeric features. Language independent word embeddings make the proposed model useful for accounts in different languages and multi-language accounts. Experiments show that our model requires less data and performs better results than previous studies about account type classification. The second problem is to personality type prediction of individual accounts. Ground truth data comes from Big Five Personality tests and Twitter accounts of a group of people. This time model also contains tweets and graph features of the account. Personalities are classified in terms of their Big Five Test scores, and for each OCEAN feature, we train different models.

Experiment results are promising even though we study with a small set of users.

Keywords: Twitter, account classification, organization vs. individual, account metadata, big five personality, personality prediction

ÖZ

TWİTTER HESAP İÇERİĞİNDEN VE YAPISINDAN KULLANICI TÜRLERİNİN BELİRLENMESİ

Gürlek, Mesut
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Ismail Hakkı Toroslu

Mart 2021 , 58 sayfa

İnsanlar sosyal medya platformlarını her geçen gün daha fazla kullanıyor ve böylece bu platformlar zengin içerikleriyle akadamik çalışmalar için daha uygun bir ortam haline geliyorlar. Twitter, en büyük ve en çok kullanılan sosyal medya platformlarından biridir ve birçok çalışma, sosyal medya araştırması için Twitter'a odaklanmaktadır. Bu tezde, Twitter kullanıcı hesaplarının türlerini, hesap meta verilerine, hesap içeriğine ve hesabın yapısal durumuna göre belirlemek için yöntemler öneriyoruz. İlk sorunumuz, yalnızca meta verileri kullanarak organizasyon ve bireysel hesap türlerini sınıflandırmaktır. Veri toplama ve analiz süreci ile ilgili detaylar verdikten sonra metinleri temizliyor ve sayı özelliklerini normalleştiriyoruz. Önerilen model, metinsel özellikler için LSTM ve sayısal özellikler için MLP içerir. Dilden bağımsız kelime vektörleri oluşturmak, önerilen modeli, farklı dillerdeki hesaplar ve çok dilli hesaplar için kullanışlı hale getirir. Deneyler, modelimizin tahminlemesi için daha az veri gerektirdiğini ve hesap türü sınıflandırmasıyla ilgili önceki çalışmalardan daha iyi sonuçlar verdiğini göstermektedir. İkinci sorun, bireysel hesapların kişilik tipi tahminidir. Temel referans verileri, Beş Büyük Kişilik testlerinden ve bir grup insanın

Twitter hesap bilgilerinden gelmektedir. Bu sefer model, ayrıca hesabın tweetlerini ve graf özelliklerini de içerir. Kişilikler, Beş Büyük Test puanlarına göre sınıflandırılır ve her OCEAN özelliği için farklı modeller eğitilir. Küçük bir kullanıcı grubuyla çalışmamıza rağmen deney sonuçları umut vericidir.

Anahtar Kelimeler: Twitter, hesap sınıflandırma, organizasyon vs. kişisel, hesap meta veri, büyük beş kişilik, kişilik tahmini

To my parents, brother and dearest wife.

ACKNOWLEDGMENTS

During the thesis writing process, there are several people I am grateful for their help. I would like to extend my deepest gratitude to my supervisor İsmail Hakkı Toroslu; completion and the success of my dissertation would not have been possible without his experience and guidance. He supported me and became very positive in the challenging times. Besides, I appreciate Yusuf Mücahit Çetinkaya for the suggestions and advice about my research.

I would like to thank my former colleagues Dr. Çağrı Toraman and Berkay for their precious ideas about my research study. I am also grateful to dear friends Mustafa, Ali, Ahmet, Enes, Ömer, Furkan, Oğuzhan for their support and understanding.

Finally and most importantly, I am deeply indebted to my parents, my brother and my dear wife. They always motivated and encouraged me throughout my master's and thesis writing process. It would be meaningless without their presence in my life to complete this study.

This study is supported by TUBITAK BIDEB 2210-A General Domestic Master's Scholarship Program, and I especially thank TUBITAK BIDEB for the support and scholarship. Besides, this research has been partially supported by TUBITAK with Project No 5180003.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

OCEAN	Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism
MLP	Multi Layer Perceptron
LSTM	Long-Short Term Memory
BERT	Bidirectional Encoder Representations for Transformers
URL	Uniform Resource Locator
BoW	Bag of Words
BPE	Byte-Pair Encoding

CHAPTER 1

INTRODUCTION

Social media became an essential part of life, not only for humans but also for organizations and establishments. Nowadays, the world is in the era of connectivity and accessibility. Fast and easy access to knowledge, sharing information, and connecting the world became fundamental needs for society. Several social media platforms emerged to meet these demands, and one of the biggest of such platforms is Twitter. Besides, Twitter is a micro-blogging system that provides users an environment to share relatively short content. Twitter also contributes a wide variety of research opportunities in various areas with the data produced in it.

While the internet access opportunities and infrastructures are enhancing, user accounts in social media platforms are also increasing dramatically. Users use these opportunities to be aware of the world and socialize. By the simplicity and ease of use of the platform, Twitter becomes a widespread social media platform among users. Since users on Twitter easily expresses their ideas, share their feelings, and expose their emotions, Twitter accounts mostly reflect real-life user personality behaviors. Hence Twitter provides rich content for various aspects of social media analysis studies, such as text mining, graph structure, or statistical analysis.

1.1 Motivation and Problem Definition

Twitter is a platform with a daily active user count of 186 million and a total number of 340 million users, and the number of tweets shared on it is over 500 million per day [5]. Hosting such a great interaction, Twitter provides researchers with a wide range of study areas. Especially in the platform, where personal thoughts are shared

through tweets, and there is an active follow-up mechanism, one of the most focused research topics is to deduce the behavior, character, and personality traits from the tweets shared and relationships between them. Data mining, graph algorithms, and deep learning algorithms are the leading techniques used for this purpose. Hence, these algorithms provide prediction, grouping, or correlation results based on certain users' characteristics. Although there are many different techniques, we need first to consult psychology to understand people's behavior and character of people. The big five personality test scores personal characters under five different personality traits in psychology and provides the ground truths for studies.

The main interest of this thesis is to study user behaviors and characters from corresponding Twitter accounts. Twitter contains a tremendous amount of data; however, we aim to predict users' traits with a minimum amount of data and reduce computation requirements and physical resources. Throughout this challenge, we have extracted graph features of user nodes, focused on user metadata, and applied natural language processing techniques on tweet texts. Big Five Personality Trait surveys of the base group of users from the study [6] provided ground truth personality traits. This survey evaluates a person in the following traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. These traits are named OCEAN, in short. To determine the personality trait of users, we use both the textual content of the tweets of the users and the network structure corresponding to the followers and the followees of that person. We grow the subgraph starting with the person, including one or two levels of her/his neighbors. Some of these neighbors are individual accounts, and some of them are organizational accounts. This information can also be useful in determining the personality trait of the person. That is why we needed to determine the type (individual vs. organization) of Twitter accounts, and we used this information in our learning process.

Furthermore, the concept of personality trait is valid for individuals. Our model has been trained with the results of the Big 5 Personality questionnaires taken by volunteers having Twitter accounts. The trained model can be applied to only individuals with Twitter accounts. Therefore, we should determine if the account belongs to an individual or an organization. If the account is determined as an individual account, then that person's personality trait is determined with our model.

1.2 Proposed Methods and Models

In this thesis, we have dealt with two different problems. The main problem is to predict Big Five Personality Traits using Twitter graph node features and tweet texts. The second problem is during the graph construction, eliminating non-individual accounts because we are only interested in the people's characters. Therefore a sub-problem, Individual vs. Organization Twitter account classification is defined. We faced with sub-problem first and then studied for the main problem.

Individual vs. Organization Twitter Account Classification is a preprocessing step for graph construction, but the problem itself introduces a set of methodologies and model building phases. Since this classification is needed for only graph nodes, we decided to use only account metadata, which contains name, bio, location, and url attributes as text features and tweets, following, followers, media, and likes attributes as numeric features. There is no tweet data that simplifies the model training and classification process. For the text fields, we applied data cleaning, language-independent tokenization, and word embedding generation. On the other hand, we normalized the values for the numeric features and combined them into a feature vector representation. Numerical features are fed into an MLP component, and text embeddings are fed into an LSTM component. Outputs of the components are combined together and fed into a dense layer. After the dense layer, a sigmoid function returns a probability of account type.

User Character Prediction based on Big Five Personality Traits contains more phases than the previous problem because this time dataset additionally contains graph node features and tweet texts. Besides the metadata features, the last 20 tweet texts of users are cleaned and word embedding vectors are extracted. In addition to the previous step, we fetched the node's first-level followers and followings to generate graph features of the node. For each node we calculated mean values of follower and following accounts' numeric attributes separately. These calculations are written as features to the base user nodes. All numeric features are converted into one vector and fed into MLP; also, tweet word embeddings are fed into the LSTM layer. Remaining process for the model training same as in the previous problem. Outputs of MLP and LSTM layers combined with a dense layer and a sigmoid returns the probability of binary

classification results.

1.3 Contributions and Novelties

In two different problems, we have different sets of contributions. Our contributions with Individual vs. Organization Twitter account classification study are as follows:

- Previous studies use tweet text originally or extract features from tweets of users. However, we propose a classification method that requires as possible as a small set of metadata and tweets are not used. Therefore data collection step is simplified; with a simple model, we give accurate results.
- Since our textual metadata is encoded into vectors by using a multilingual Sentencepiece model, the solution is language independent.
- We provide an end to end classification process that starts from data fetching to prediction. This study can be ready to be used for twitter graph preprocessing to eliminate organizational accounts or can be adapted for any study that requires individual vs. organization account separation.

In the second part of the thesis, extracting graph features besides the tweet embeddings and account metadata provided a solid contribution to the domain and the literature. The ensemble of three different types of features into a single deep learning model distinguishes this study from previous ones.

1.4 The Outline of the Thesis

The following chapters are organized as follows:

- Chapter 2 examines the related studies for both problems in the literature.
- Chapter 3 provides background information and detailed explanations about methods, theories, and tools used in this study.

- Chapter 4 explains the Individual vs. Organization Account Classification study phases, which are data collection, preprocessing, model training, experiments, and evaluation.
- Chapter 5 describes the organization account elimination using the model in Chapter 4, and gives details about Big 5 Personality Traits Prediction model training and experiments.
- Chapter 6 discusses the results, gives a brief conclusion, and presents ideas about future improvements.

CHAPTER 2

RELATED WORK

There is a variety of works concentrating on the Twitter account classification in terms of different aspects. Several features can be extracted for classification. Mainly textual and statistical properties of account information and user tweets are used in the previous studies in various combinations. In [7], authors classify the Twitter users as individual vs. organization using textual content of the tweets for event recognition in Spanish and English. The studies in [8] and [3] use account information along with user tweets to distinguish the organization and individual accounts. Both approaches have several drawbacks, such as language dependency, exhausting data collection, and processing stages due to noisy and unstructured text data. To overcome language dependency in textual data, studies in [9] and [10] aim to extract statistical features from texts such as the number of favorites, number of retweets, tweet length, number of hashtags, and emoticons in tweets. Although these works overcome the language dependency problem, they still need to collect tweet data and require complex preprocessing.

When the purpose of the account classification is eliminating non-individual accounts from the social network graph, then using tweet data would be redundant because some social network graph problems do not need tweets but only focus on the graph structure. Other approaches use time distributions of the tweets [11], or the post frequencies of the users [12] to avoid processing tweet texts. However, these approaches also have similar disadvantages as the previous studies since, even though the texts are not processed, they still need to access the tweet data. An alternative way is to use the joint representation of profile and network information of the account [13]. In [14], the authors construct a graph to find organization accounts based on their struc-

tural and behavioral factors.Building graph brings the computational cost of network structure; therefore, eliminating non-individual accounts without graph construction is less costly.

The study in [15] proposes a demographic inference model that uses profile image and account metadata for three attributes: gender, age, and organization indicator (individual vs. organization). The model uses Long Short Term Memory (LSTM) architecture [16] for textual features and DenseNet [17] for the profile picture. The model is trained with the data set crawled by the authors, which is a heuristically-identified, and annotated data set with the size of 23.86M for individual vs. organization classification task.

Categorizing accounts might help to convey the intended message to the target audience. In this sense, some studies classify users based on their interest [18], [19], [20], personality [6], [21], [22], political orientation [23], [24], [25], gender [26], [27], and so on.

Predicting the user personality traits from social media accounts is another popular and one of widely studied areas. [28] analyses user personality according to their publicly available Twitter profile image choices. [29] fuses the information of Twitter and Instagram, hence proposes a method that combines text, image, and metadata features. Both studies also use the Big Five Tests of users as ground truths for personality traits.

Big five personality test methodology is often a standard personality test used in many academic studies in various fields. [30] is an interdisciplinary study that focuses on psycho-demographic profiles of Facebook users. Their dataset contains detailed demographic information about the user, Facebook likes, and big five personality tests. They work on huge data and predict users' personalities using Facebook likes. [1] is one of the early studies about the big five personality test. In the study, regression is applied to the data, and they predict Twitter user personality scores.

Finding structural similarity between graph nodes is another important topic for social media studies, which leads to correlating user behaviors in the platform. [31] and [32] present methodologies for learning node embeddings which allows to compare

structural identities of nodes. Using pure structures gives a big picture of the users' relations and behaviors in the social media graph.

CHAPTER 3

PRELIMINARIES

In this chapter, there are techniques and tools used during data collection, pre-processing, data analysis, and deep learning model construction phases of the problems. Background information for the concepts and tools are provided in this chapter to give the reader a broad understanding.

3.1 Big 5 Personality Traits

Personality traits are basic curiosity and mystery for researchers about human nature. Therefore, several trait theories are developed, such as Raymond Cattell's sixteen personality factors and Hans Eysenck's three-factor theory. Nevertheless, Eysenck's theory scope was very limited, and Cattell's factors were very complicated. [33]. Therefore, in the 1980s, a more suitable five-factor model, Big Five Personality Traits, shortly named OCEAN, was presented. Personality traits are scored between 0-100 in five categories which are Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

Each factor represents a person's tendency in different traits, as shown in Table 3.1. The person who has high scores in ***O****pennes* has a tendency to be more creative and imaginative; conversely, the person with low scores tend to be conservative. High scores in ***C****onscientiousness* shows that person has discipline in the life routine and very well organized; low scores shows spontaneous and disorganized routine. High scores in ***E****xtraversion* means a more self-confident, positive, and outgoing person; on the other hand low score means detached from society and solitary personal traits. High scores in ***A****greeableness* presents cooperative, helpful, and trusting characteris-

tics, while low scores present competitive and argumentative characteristics. Lastly, high scores in *Neuroticism* in test results is a sign of behaviors with emotionally unstable and prone to stress; however, low scores is a sign of emotionally stable and positive behaviors.

Table 3.1: Corresponding characteristics to the five factor of *OCEAN* [1].

Trait	**High Scores**	**Low Score**
Openness	Imaginative	Conventional
Conscientiousness	Organized	Spontaneous
Extraversion	Outgoing	Solitary
Agreeableness	Trusting	Competitive
Neuroticism	Prone to Stress and Worry	Emotionally Stable

3.2 Twint

Twint [34] is an open-source project for Twitter scraping which has no relation with Twitter API. Therefore, it has no restrictions that Twitter API imposes. It provides to collect user profile metadata, tweets, following, and follower lists, and many other fetching configurations. Another advantage over Twitter API is it does not require an account or sign up; all data can be fetched anonymously. Twint has a CLI command list and python library with simple installation and usage. Basic commands and their functionality can be found below:

- *twint -u username* : Collects all tweets of a user.
- *twint -u username –user-full* : Collects metadata information of a user.
- *twint -u username –since "2015-12-20 20:30:15"* : Collects all tweets since a spesific time.
- *twint -u username –email –phone* : Collects all tweets which contains phone or email.

3.3 Word Embedding

Word Embedding is an essential concept for Natural Language Processing problems that converts words into vector representations in a predefined vector space. There is a need to convert documents into computer-readable representations to make NLP computations easily. Storing methodology of the words decides the quality of representation of words so the document. Initial representations focus only on the existence of the words such as One-Hot Encoding and frequency of the words in the document such as Bag of Words. These implementations do not contain any semantic information. One-Hot Encoding represents a document or text as a sparse matrix that stores the existence of words and the order of the words in the document. Table 3.2 shows the One-Hot Encoding of the word corpus of the example sentence and sentence embedding vectors.

Example sentence : *I like meat and I like potato*

BoW representation of the same sentence is provided below:

BoW Embedding : *"I": 2, "like": 2, "meat": 1, "and": 1, "potato": 1*

Table 3.2: One-hot encoding. Left table shows the encoding and right table show the sentence word one-hot embeddings.

I	=>	1	0	0	0	0
like	=>	0	1	0	0	0
meat	=>	0	0	1	0	0
and	=>	0	0	0	1	0
potato	=>	0	0	0	0	1

I	1	0	0	0	0
like	0	1	0	0	0
meat	0	0	1	0	0
and	0	0	0	1	0
I	1	0	0	0	0
like	0	1	0	0	0
potato	0	0	0	0	1

Contemporary embedding techniques provides that words with similar meaning converted into similar vector representations [35]. Each word has one corresponding vector, and word vectors are combined to represent a text. The main contribution to

the word embedding concept is to representing words in a common space and exposing their semantic distance. This representation also enabled neural networks to learn semantic representations of the word sequences better. There are several word embedding algorithms and pre-trained models in the domain. Our requirements met with language-independent MultiBPEmb [36] word embedding vectors that support 275 languages.

3.3.1 MultiBPEmb

MultiBPEmb is a collection of multilingual subword segmentation models and pre-trained subword embeddings[36]. It can tokenize concatenated words in 275 languages and provide 300-dimensional subword embeddings for different Byte-Pair Encoding [37] vocabulary sizes. MultiBPEmb uses SentencePiece [38] to learn BPE subword segmentation models and GloVe [2] to train subword embeddings. SentencePiece provides the tokenization, and each token is represented with 300-dimensional embedding vector representations in the multilingual model, while single language models have several vector dimension options starting from 25 to 300.

3.3.1.1 SentencePiece

SentencePiece is a language-independent subword tokenizer, and detokenizer adopts BPE [37] algorithm for a subword-unit generation. The main advantages of SentencePiece is as follows:

- SentencePiece takes sentences as a set of Unicode characters; hence it is language independent.
- Provides fast segmentation with 50 thousand sentences in a second.
- It is consistent and self-sufficient since it provides the same tokenization and detokenization results all the time.
- It is easy to use and provides 275 language models separately.

3.3.1.2 Byte-Pair Encoding

Byte-Pair Encoding originally is a data compression algorithm. It iterates throughout the bytes, and common most frequently existing bytes are replaced with a byte which does not include in data. [39]. BPE can be applied to text compression as well. The example below demonstrates simple compression logic.

Initial text: ACDFACKALACD

*Step 1) Most frequent pair is **AC**:* XDFXKALXD

*Step 2) Most frequent pair is **XD**:* YFXKALY

Nowadays, BPE is used as sub-word embedding and NMT(Neural Machine Translation) in the recent studies such as SentencePiece [38] and GPT-3 [40]. BPE's subword embedding algorithm works as follows; each character is counted as a single element, and a special character **</w>** used to emphasize that it is the end of the word:

Initial set of characters:

'l o w </w>' : 5, 'l o w e r </w>' : 2, 'n e w e s t </w>':6, 'w i d e s t </w>':3

1. iteration:

'l o w </w>' : 5, 'l o w e r </w>' : 2, 'n e w es t </w>':6, 'w i d es t </w>':3

2. iteration:

'l o w </w>' : 5, 'l o w e r </w>' : 2, 'n e w est </w>':6, 'w i d est </w>':3

3. iteration:

'l o w </w>' : 5, 'l o w e r </w>' : 2, 'n e w est</w>':6, 'w i d est</w>':3

4. iteration:

'lo w </w>' : 5, 'lo w e r </w>' : 2, 'n e w est</w>':6, 'w i d est</w>':3

4. iteration:

'low </w>' : 5, 'low e r </w>' : 2, 'n e w est</w>':6, 'w i d est</w>':3

After all merge operations:

'low</w>' : 5, 'lower</w>' : 2, 'newest</w>':6, 'widest</w>':3

Resulting dictionary: 'low', 'lower', 'newest', 'widest' [37]

One of the main functionality of the Byte-Pair Encoding algorithm is eliminating unknown words. If there is sufficient iteration, all sub-words can be discovered.

3.3.1.3 GloVe

GloVe is an algorithm that generates word embedding vectors in an unsupervised way. GloVe learns the vector representations from aggregated global word to word co-occurrence statistics extracted from a corpora. [2]. The co-occurrence statistics matrix reflects how frequently the two words occur in the corpus, and model training is done by passing through the whole matrix. Construction of the co-occurrence matrix is a bit costly because of the passing through all the corpora, but since the model training is on the co-occurrence matrix, it is much faster.

Table 3.3: Co-occurence matrix of *ice* and *steam* from the GloVe paper [2]. The ratio between statistical co-occurrence probabilities give the semantic relation.

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

The logic behind capturing semantic representation from the words, the model analyzes the ratio between words in the co-occurrence matrix. Table 3.3 is adopted from the original GloVe paper to show how the model understands semantic relation between words. It can be inferred from the table 3.3 that *ice* co-occurs with *solid* and *steam* co-occurs with *gas* which is logical in terms of real-world semantic. Both words are also frequently occurred with *water* and similarly less occurred with *fashion*. When the ratios are examined, if the ratio is close to 1, they semantically have the same distance with the same words. However, if the ratio is too high or too low than 1, they have different tendencies to occur with different words. From the ratios, we can infer that *ice* is more likely to occur with solid, conversely *steam* more likely to occur with gas.

Nearest neighbors of a GloVe vector in the vector space also provides semantically similar word vector as a result. To illustrate, nearest neighbors of *museum* are:

art, gallery, library, exhibition, museums, heritage, collection, sculpture .

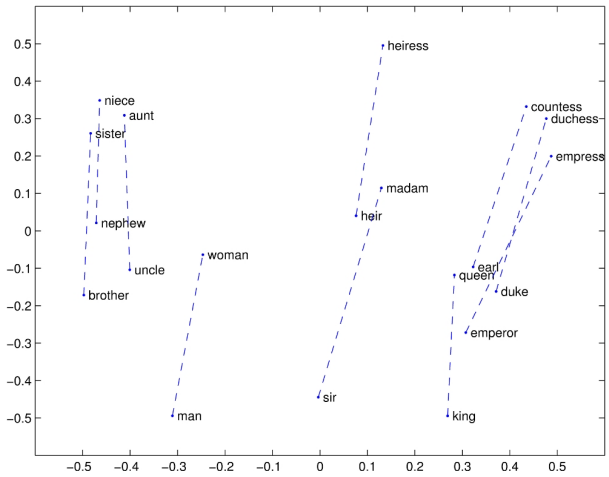


Figure 3.1: Similar Linear Relations between Semantically Close Words. Figure taken from GloVe project website [4]

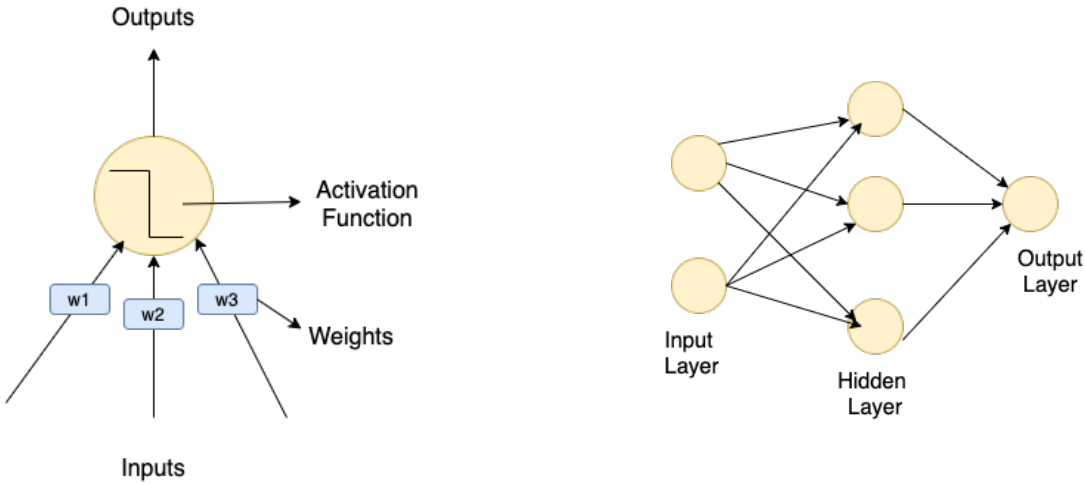
Since the similarity calculation in nearest neighbors methodology provides a single numeric that represents the relatedness among words, we may still be missing some information, and a single number cannot represent the whole relation. For example, *queen* and *king* are both person; however, their semantic meanings quietly opposite. Hence, taking vector difference between representations provides a vector that also contains a linear substructure of the relation between words.
Figure 3.1 show the similar nature between word pairs *queen-king* and *woman-man*.

3.3.2 BERT

Bidirectional Encoder Representations from Transformers (BERT) is an encoder-decoder based model used for language modeling [41]. It is developed on top of Transformer [42]. BERT is a bidirectional model where it extracts the context from the sequence of words in a document. The words are consumed by the encoder. The model generates the context vector while the words are being encoded. There are pre-trained models with different corpora that is supposed to work well for various NLP tasks.

3.4 MLP

Multi-Layer Perceptrons is a feed-forward artificial neural network. Artificial neural networks consist of neurons, which are the minimum computational unit and building blocks of the network. A perceptron is a single-layer neural network that consists of simple inputs, weights, weighted sum, activation function, and outputs as shown in Figure 3.2a. Multiple perceptron layers come together and form Multi-Layer Perceptron. MLP consist of 3 different layers: an input layer, hidden layers, and an output layer that is depicted in Figure 3.2b. If MLP contain more than one hidden layer, it becomes a deep neural network. MLP has a feed-forward architecture and used for supervised learning. Non-linearity in learning is ensured by the activation function. Feed forwards mechanism takes the input and generates the prediction. On the other hand, there is a backpropagation mechanism that takes prediction, calculates error rates, and updates the weights of the layers to predict better in the next iteration. This process continues in each iteration until model loss converges to a result.



(a) Single Perceptron

(b) Multi-Layer Perceptron

Figure 3.2: Single and Multi-Layer Perceptron Representations

3.5 LSTM

Recurrent neural networks are the neural network that takes sequential data or time-series data as input. They are good at learning a series of data, but it has some shortcomings. When the input data sequence is too long, the weight of initial data that fed into the network drastically decreases, and it will be nearly zero effect on the output. This problem is named as vanishing gradient problem. Since gradients are used to update network weights when the sequence is long, the gradient vanishes during the backpropagation, and as a result, weights are not getting updated.

$$newWeight = weight - learningRate * gradient$$

Long-Short Term Memories uses memory cells to keep useful information from previous cells and forget the information with trivial information. As can be seen from Figure 3.3, LSTM consist of cell state and the gates. Forget gate contains a sigmoid function that generates an output between 0 and 1; therefore, if the information is unnecessary, the resulting output will be close to 0, and information will be forgotten. Input gate consists of a sigmoid and a tanh function that provides the update of the input information. Cell state consists of the combination of forget gate output and input gate output. Lastly, the output gate decides what the next hidden state will become.

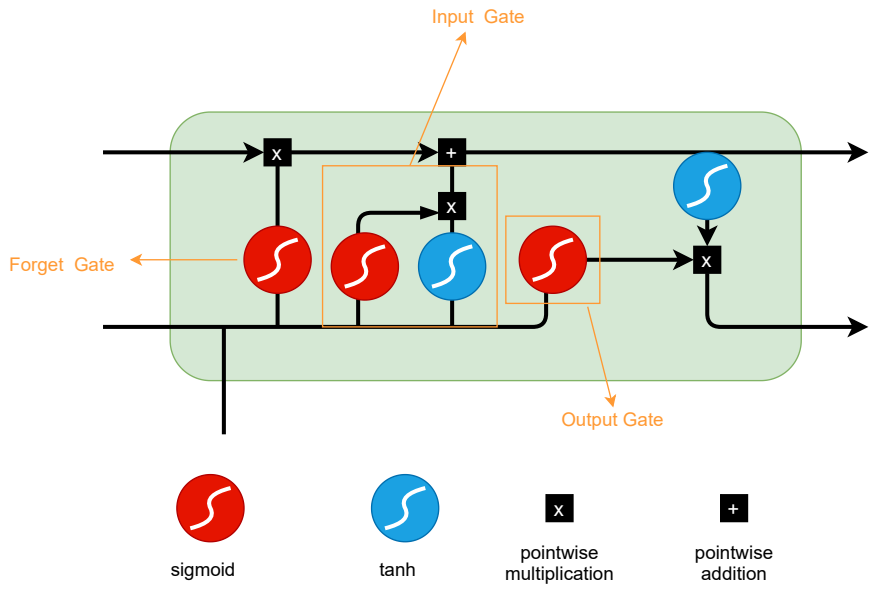


Figure 3.3: Single Long-Short Term Memory Cell

3.6 Dropout

Complex and large Neural Networks are commonly used, while tasks are getting complicated. Large networks have too many neurons, and some of them extract features better compared to others. However, when the network is getting bigger, there is a problem called overfit that emerges from iteration to iteration. To overcome this problem, different regularization methods are used during the training, such as L1 or L2 regularization, to add a penalty to the loss function. Dropout is a very effective regularization technique for neural networks. It simply drops-out some of the neurons in the layers to highlight neurons that are better at feature extraction. Dropping out means ignoring some of the neurons in each iteration. Dropout can be included in MLP or LSTM and widely used to improve the training phase. Its default value is 0.5, resulting in dropping half of the nodes in each iteration, which increases iteration count and converging time.

CHAPTER 4

INDIVIDUAL VS ORGANIZATION TWITTER ACCOUNT CLASSIFICATION

Twitter provides rich content for various aspects of social media analysis studies, such as text mining, graph structure, or statistical analysis. However, the data needs preprocessing and cleaning since personal accounts are not structured, and there are also many fake accounts, bot accounts, and corporate accounts. Annotating the dataset manually is costly and biased since annotators' decision might be subjective. Therefore, various research attempts are trying to make account type detection automatic [43]. This automated account classification is a necessary prior step to improve the quality of various social media analysis tasks such as building social media graphs or analyzing the message texts. One of the biggest challenges in building social network graphs is the complexity of the space and the time. Thus, account elimination is significant to reduce this complexity [44], [45], [46].

One of the popular research subjects that also requires tweets is the sentiment analysis [47], [48], [49], [50], which is widely used for identifying user's reviews of entities to capture the trends. Similarly, event-related Twitter posts are trendy research topics [51], [52], [53], [54]. Tweets themselves are the primary source of information for such researches. For these and in many similar studies, individual accounts are needed, and organizational accounts must be excluded.

Our study focuses on the same problem and aims to determine the organizational and individual accounts automatically. We propose a powerful approach to classify the Twitter accounts as individual vs. organization, using limited metadata of the accounts. Organizations, institutions, news or entertainment accounts, etc. are labeled as *organization*. On the other hand, *personal* accounts are the accounts that reflect the

emotion, thought, or behavior of a human, so fake user accounts can also be personal accounts.

As one of the previous studies on the same problem, in [8], for account classification, the last 200 tweets posted from a given account are analyzed. Following a similar approach, the work by [3] adopts a single tweet to feed into a character-based convolutional neural network. Unlike these two, we concentrate on only the metadata of the account. Thus, in our approach, the data collection, preprocessing, and classification stages require much less effort compared to the related methods. Furthermore, protected accounts' metadata are also publicly available; thus, it is possible to classify the protected accounts as well. The proposed solution is publicly available as a google colab script.

Today, many Twitter accounts contain texts written in multiple languages and dialects. Therefore, multi-lingual support is needed for the account classifiers that analyze message content. Our proposed model does not depend on a single language. It encodes the textual properties of metadata using Byte-Pair Encoding [37]. Numeric features support the textual features and increase accuracy slightly.

Due to the nature of Twitter use, the number of individual and organization accounts are inherently unbalanced, such that the individual accounts constitute the majority. To overcome this deficiency for model construction, we have followed the procedure used in the studies [8], [3], where the dataset is arranged as balanced dataset, unbalanced dataset, and full dataset.

4.1 Problem Definition

Twitter is a widely used social media platform for conducting research. Some studies focus on individual accounts while others examine organizational accounts. Thus, separation is necessary according to the domain and the scope of the study. Since Twitter accounts may contain a considerable amount of data, it is hard to collect information when the user count increases. Therefore, if the proposed solution requires less data collection effort makes the process easier. Also, datasets may contain accounts with different languages, or multi-language accounts might exist. Hence, the

proposed method should handle multi-language account problems. We defined an Individual vs. Organization classification problem for Twitter accounts in light of this background information.

4.2 Method

This section provides information about collecting data, having insight analysis, preprocessing and building classification pipeline.

4.2.1 Data Collection

The data collection step is less laborsome in our method than those in similar studies. Previous studies given in [8] and [3] provide publicly available data sets, Humanizr and Demographer. In the repository[1], they provide the Demographer dataset as a text file containing Twitter account ids and labels for each account. The authors also provide Humanizr dataset in the same format, which they collected in 2018. Using Twitter account id, we have fetched the profile metadata of Twitter accounts using open source Twitter scraping tool 'Twint' [34]. Since we are using only the profile information, protected accounts are also included in the dataset. We fetched 17,790 Twitter account profiles from Humanizr data set and 214,236 accounts from Demographer Table 4.1 provides the number of successfully fetched accounts in the datasets. The difference between the original count and the collected count stems from closed or suspended Twitter accounts. Table 4.2 shows the profile metadata for a sample account that consists of 18 attributes. Among these attributes, four textual and five numeric attributes are used for model construction.

[1] [3] provides a repository for data and the code of the study. https://bitbucket.org/mdredze/demographer/src/master/data/.

Table 4.1: Humanizr and Demographer dataset statistics. Original numbers are taken from study in [3]

Dataset Name	Original Accounts	Collected Accounts	Collection Percent
Humanizr	18,922	17,790	94%
Demographer	227,277	214,236	94%

Table 4.2: Twint provides 18 different attributes for user profiles. The attributes used in this study are marked in bold.

Profile Attributes	Description	Type
id	Unique identifier of user	integer
name	Name of the user	text
username	Unique screen name	text
bio	Description of the account	text
location	User-defined location	text
url	A url provided by user	text
join date	UTC date of account creation	text
join time	Time of account creation	text
tweets	Number of tweet of the account	integer
following	Number of users this account following	integer
followers	Number of users follows this account	integer
likes	Number of tweets user has liked	integer
media	Number of media in the tweets	integer
private	True when account is protected	boolean
verified	True when user has verified account	boolean
avatar	URL pointing to the user's profile image	text

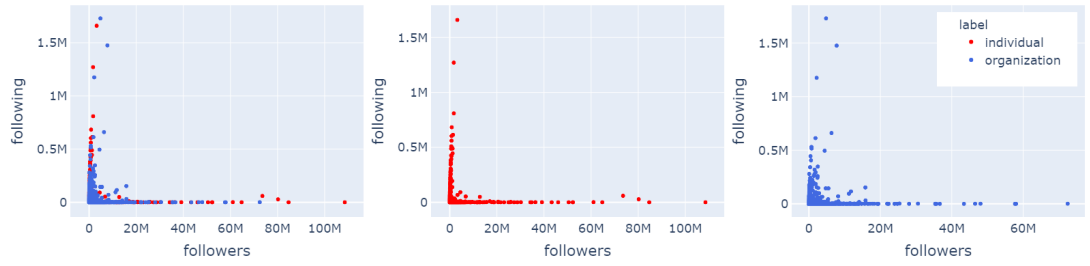
4.2.2 Data Analysis

The collected Humanizr dataset consists of 17,790 user accounts, in which 16,012 of them are labeled as individuals, and 1,778 of them are labeled as organizations. In the Demographer dataset, there are 214,236 accounts in which 185,224 are labeled as individuals, and 29,012 are labeled as an organization. Since the Demographer dataset contains the Humanizr dataset, the figures are based on the Demographer dataset in the following sections. Dataset evaluation and the comments of Figures are not only based on the visualizations but also specific data points in the figures are fetched as raw and investigated by hand. [2]

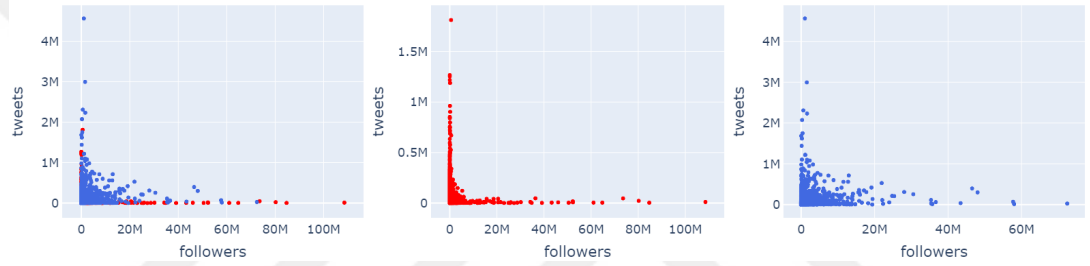
[2] Source codes, the dataset, and the interactive versions of the given figures used in this study are publicly available at https://github.com/tweetpie/twitter-account-classification.

Figure 4.1: Word cloud of Demographer dataset, containing top 250 occurred words in individual (above) and organization (below) accounts.

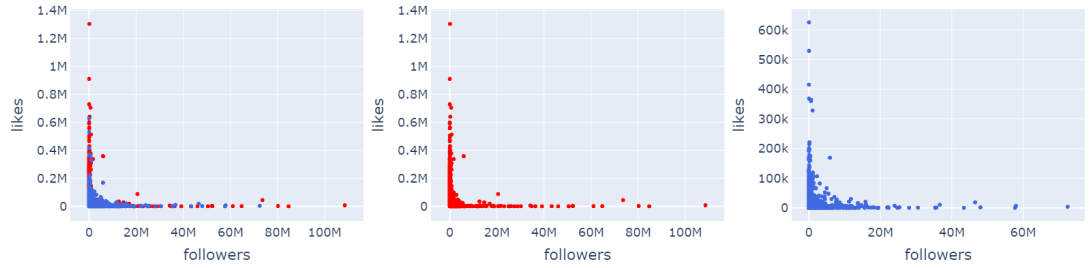
We selected *name*, *bio*, *location*, and *url* attributes as text features and *tweets*, *following*, *followers*, *media*, and *likes* attributes as numeric features for the classification. Figure 4.1 shows that there is a high overlap between used words for each account type.



(a)



(b)



(c)

Figure 4.2: The distribution of accounts in the Demographer dataset comparing the follower count vs. followee, tweet, and like counts.

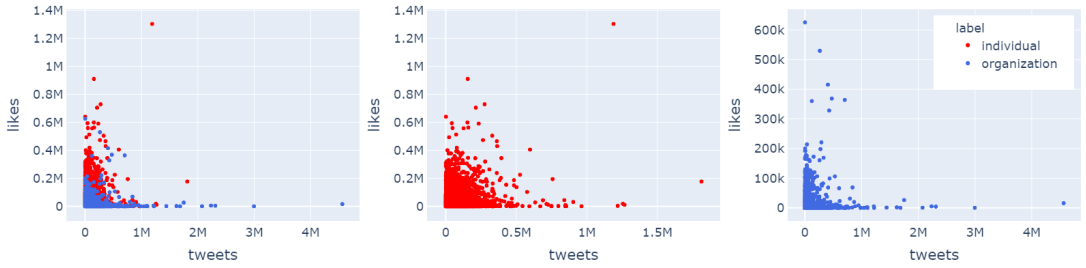
Figure 4.2a visualizes the relationship between the number of followers and followee per account. The individual accounts that have high number of followees, but few number of followers try to increase their visibility by using the follow-back strategy. The majority of this group is entrepreneurs. In contrast, accounts with high number

of followers and fewer number of followees are usually owned by the people who are recognized by the crowd with their success in their expertise areas. Similarly, organization accounts of large companies have high number of followers and few number of followees. Newspapers, news agencies, social media companies, and well-known brands of the entertainment industry are such examples. Non-profit or charity foundations do not have concerns over following-back accounts since they aim to increase their visibility to reach out to the people.

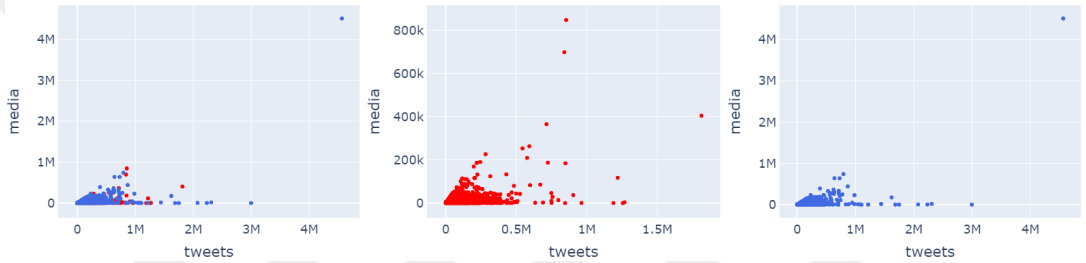
The individual accounts in Figure 4.2b that have more than 500K tweets do not have many followers. Still, they continue to tweet consistently. Similar to Figure 4.2a, high-followers less-tweets points are celebrity accounts. They do not aim to be active with posting tweets since their reputation in the real world is high enough. The organization accounts having high-tweets are the support accounts of companies. Customers tweet via mentioning their account names in their tweets while expressing problems with the product or service. By nature, these accounts do not have that many followers. The organization accounts with so many followers tend to tweet less. The pioneer enterprises adopt this behavior mostly. Organizations tweet more than the individuals independently of their follower counts. Since those accounts are mostly managed by professionals, they have a consistent tweet sharing behavior compared to the individuals.

Figure 4.2c shows that individuals having fewer followers use Twitter as regular users where follow other accounts and trend topics consistently. This situation makes their like count higher than the others. Individuals with high followers tend not to like others' tweets. The reason might be that they feel uncomfortable since their likes are exposed to the mass. The main aim of the organizations with few followers is to increase their visibility by liking other tweets. The accounts following this strategy are mostly small or medium-sized enterprises and organizations trying to build a community. In contrast to 4.2a, charities and non-profit organizations take up less space in the group of organizations with high number of likes. Twitter accounts of big companies have a similar attitude to be neutral, considering the company interests. Organizations like tweets as half as individuals do on average.

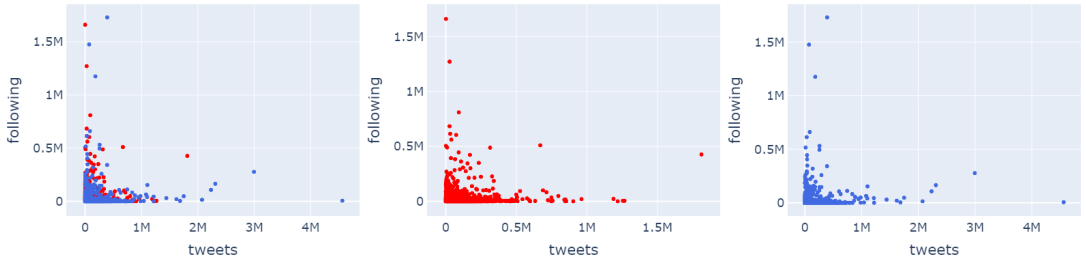
As shown in Figure 4.3a, individuals have more likes compared to the organizations



(a)



(b)



(c)

Figure 4.3: The distribution of accounts in the Demographer dataset comparing tweet count vs. like, media, and following counts.

in general. Celebrity accounts are close to the origin where they usually tweet and like less. At some point, like count becomes acutely fewer while tweet count is increasing. Support accounts have more tweets but fewer likes as expected. The balance between tweets and likes is more consistent for individuals than organizations. The organizations with few number of tweets but high number of likes usually do not post original tweets but retweet or like other accounts' content.

The diversity of using media in the tweets of individuals is higher than the organizations. The organization accounts that have media in almost all of their tweets are mostly propaganda or ad accounts. The organization accounts with high number of tweets and less number of media in Figure 4.3b are the support accounts. They usually reply to the clients that are asking for help with plain text.

Individuals following so many accounts do not frequently tweet, as displayed in Figure 4.3c. They aim to gain followers with the follow-back strategy. In contrast, the ones who tweet a lot do not follow that many users. These two groups pursue different strategies to become famous. Nonetheless, the accounts having less than 500K tweets have more consistent number of followees. Similar behavior can be seen for the organization accounts as well.

The Figures 4.2 and Table 4.3 clearly demonstrate that numeric features for organization and individual accounts do not give explicit clues to distinguish them. The accounts having similar goals are gathered together, as explained in this section. However, the numeric features do not have distinctive information for organization vs. individual classification. Therefore, we used their textual features accompanied by the numeric to classify.

Table 4.3: The statistics for the word length in textual features in Demographer dataset.

	Average	Standard Deviation	Max
Bio	11	5.97	105
Name	1.97	0.74	33
Location	1.55	1.96	34
Url	2.48	1.53	48

Table 4.4: The statistics of the numerical features in Demographer dataset.

	Average	Standard Deviation	Max
Tweets	22,982.40	67,810.64	2,706,830
Following	3,007.94	32,085.57	1,196,529
Followers	345,570.94	2,718,346.26	108,539,399
Likes	8,790.79	28,226.66	617,782
Media	5,256.03	25,707.56	752,000

Table 4.3 and Table 4.4 show the statistical summary of the features. For the textual features, average, standard deviation, and max word count are investigated to apply correct padding in the preprocessing step. The same statistical investigation is applied to numerical features, but outliers that have extremely high values for the features affect the overall average, standard deviation, and max values, as can be seen in Table 4.4. Therefore, these statistics do not give useful insights for the numerical features in the preprocessing phase.

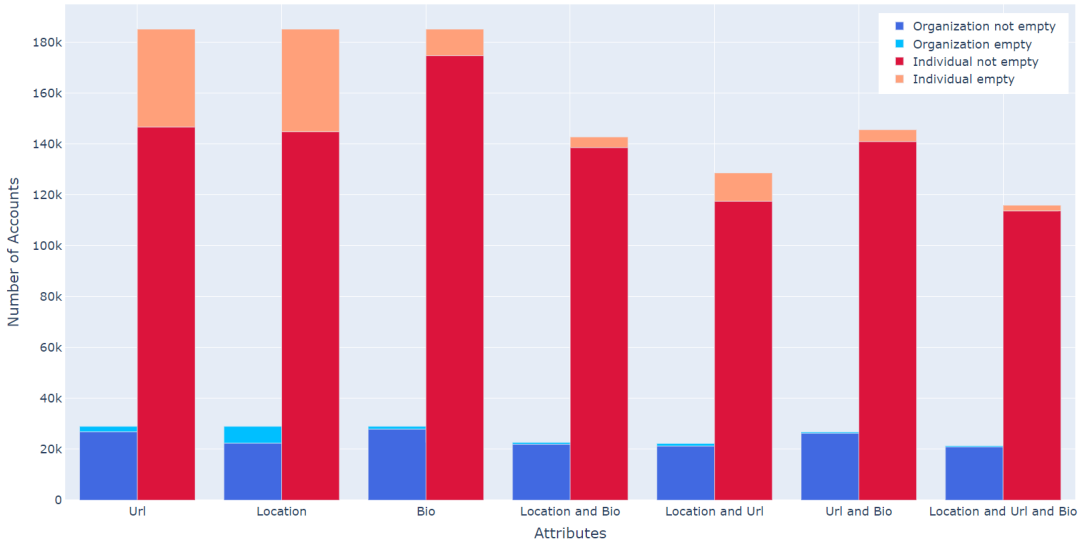


Figure 4.4: Demographer Dataset Accounts Analysis on Existence of Url, Location and Bio Attributes

Demographer Dataset includes Humanizr Dataset accounts, thus, data analysis section focuses on Demographer. We considered that the URL, location, and bio fields might differ according to account type, and hence they can provide a clue for the account classification task. Figure 4.4 shows the number of accounts in terms of their class that does not have these attributes full. As given in Figure 4.4, the percentage of missing attributes in terms of classes is very close for bio and location. The missing URL attribute percentage among organization accounts is 7% but among individual accounts is 21%. These statistics might indicate that if an URL attribute is not specified for an account, it is more likely that it is an individual account than an organization.

4.2.3 Data Preprocessing

In the preprocessing step, we have applied basic data cleaning techniques. The numerical values in the textual data are converted into textual equivalents to facilitate the model training. We have removed all punctuation and turned the whole text into lowercase. Stopwords for the languages used in the dataset are removed from the account texts. Normalization has been applied for the numeric attributes, and the values

are scaled into the range [0-1].

There are accounts in various languages, making it challenging to find a standard embedding technique to prepare texts as input to the neural network. Additionally, since Twitter users do not have to focus on grammar rules or language structure in their account information, we observed that some words are concatenated in these metadata texts. Therefore, pre-trained models for word embeddings, such as GloVe [2] or word2vec [55], could not work directly for our dataset because they do not include these concatenated words in their corpora. We used Sentencepiece [38] to extract the tokens from compound words.

We have also observed that some sentences include words from different languages together. Unknown words can be represented with special tokens like <unk> in the popular word embeddings like word2vec or GloVe, but it is not a viable solution. There are different approaches to solve this problem, such as subword segmentation, which reconstructs a word's meaning from its parts. Byte-Pair Encoding [37] is an unsupervised subword segmentation method that starts with a sequence of symbols and iteratively merges the most frequent symbols to generate a new symbol. The origin of BPE comes from [39], and it is presented as a data compression algorithm that works by replacing with a byte that does not appear in that data as can be seen in Figure 4.5.

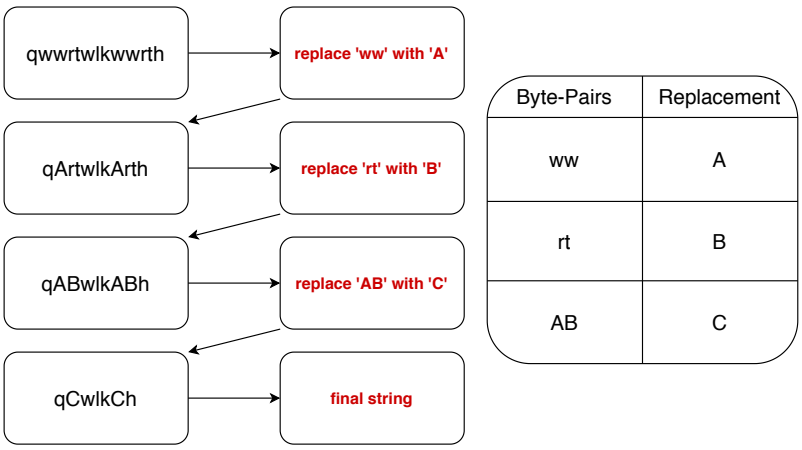


Figure 4.5: Original Byte-Pair Compression Algorithm

The study in [37] adopts this algorithm and modifies it slightly to provide a solution

to the subword tokenization problem. The main difference from the original implementation is combining the most recurring characters or words instead of replacing them with another byte for compression. For example, in English corpora, the most recurring characters ‘h’ and ‘e’ are combined into ‘he’, ‘t’ and ‘h’ are combined into ‘th’, then ‘t’ and ‘he’ are combined into ‘the’ to generate a new word.

We have used MultiBPEmb [36], which is a collection of multilingual subword segmentation models and pre-trained subword embeddings. It can tokenize concatenated words in 275 languages and provide 300-dimensional subword embeddings for different Byte-Pair Encoding vocabulary sizes. MultiBPEmb uses SentencePiece to learn BPE subword segmentation models and GloVe to train subword embeddings. Table 4.5 shows the tokenization results of several examples. Each token has a 300-dimensional embedding vector representation.

Table 4.5: Sentencepiece tokenization results for some examples.

Language	Text	Tokens
en	defenceandtechnology	['_defence', 'and', 'technology']
	ventureforamerica	['_venture', 'for', 'america']
	lifeinpleasantville	['_life', 'in', 'pleasant', 'ville']
	peteforamerica	['_pete', 'for', 'america']
de	diewahrheit	['_die', 'wahr', 'heit']
iw	רעצדוע	['רעצ', 'דוע_']
ru	степеньпромышленногопроизводства	['_степень', 'промышлен', 'ного', 'произ', 'водства']
es	bailarconmusica	['_bailar', 'con', 'musica']
fr	chargédecommunication	['_chargé', 'de', 'communication']
it	nontihoamore	['_non', 'ti', 'ho', 'amore']

4.2.4 Classification

Our proposed model consists of two parallel models: multi-layer perceptron (MLP) for numerical features and bi-directional long short-term memory (LSTM) for textual features. Figure 4.6 shows the architecture of the model in details.

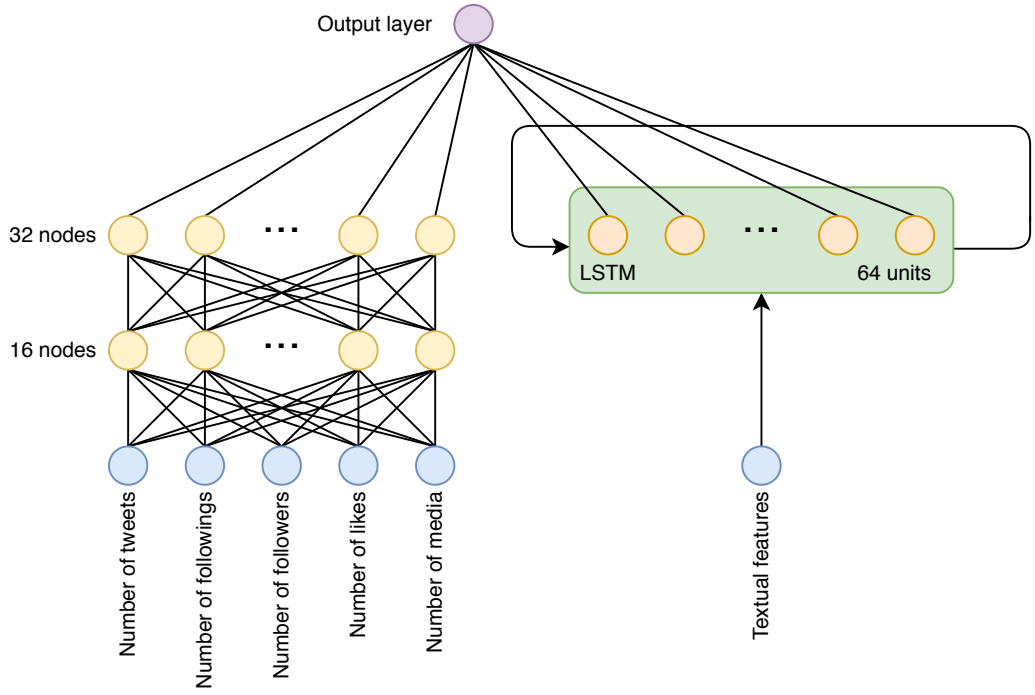


Figure 4.6: Multi-layer perceptron (MLP) for numeric features and long short-term memory (LSTM) for textual features are used. Textual features are fed into the LSTM within a sequence. The last layers of the two models are concatenated before connecting it to the output layer. While hidden layers have ReLU as activation function, the output layer uses sigmoid.

MLP consists of multiple perceptrons echeloned into multiple layers. Single perceptron takes numeric inputs with the corresponding weights to emit these values' summation to the next layer's neurons. In order to add non-linearity to the model, an activation function can be applied to the sum. The signal emitted from the output layer is used to predict the class of the given sample. The architecture may vary with the number of layers, neurons, activation function applied to the neurons.

We feed the MLP model with five numeric features: the number of tweets, followings, followers, likes, and media. It has two hidden layers with 16 and 32 hidden nodes correspondingly. Both layers have rectified linear unit (ReLU) activation function at the end. Bi-directional LSTM has 64 units inside, and 300-dimensional embedding vectors are fed. We set the maximum sequence length to 64. A dropout layer is connected to LSTM for regularization. At each iteration, random nodes are removed

to prevent the model from over-fitting. The output of the last MLP and LSTM layers are concatenated before feeding it to the output layer. The activation function of the output layer is the sigmoid function for binary classification. The loss function for our model is binary cross-entropy. The model uses the Adam optimizer for adjusting weights. We used a decaying learning rate, starting from 10^{-3}. We selected the batch size as 128 for the training process.

4.3 Experiments and Evaluation

We have implemented our model using the Tensorflow library in Python, which is successfully applied in similar tasks [56]. The experiments are conducted under 7 fold cross-validation. We adopted four evaluation metrics for evaluating the classification performance with weighted average: accuracy, precision, recall, and F-measure. Three different study [8], [3], [15] focus on the same classification problem with different approaches. Their evaluation also differs, while [8] and [15] provide recall and accuracy, [3] only reports overall accuracy. Since we have four different evaluation criteria that encapsulate previous techniques, we take these studies as baselines and compare them with our results.

Table 4.6 shows the evaluation results of the previous studies in comparison to our proposed model. We called the model names same as the data they are trained on. In the table, baseline models are compared with the proposed model in a paired manner. Since we did not test baseline models, the results of Humanizr [8], Demographer [3] and M3 [15] models are given as reported in the original papers. However, each model is applied on different versions of the datasets, such as the Humanizr model applied on the balanced, unbalanced, and full version of the Humanizr dataset. On the other hand, the Demographer model is applied to the balanced and full version of the Demographer dataset. The balanced version of the data set includes an equal number of individual and organization accounts. The number of accounts is accomplished by undersampling the majority class. They train and test on the balanced dataset, which leads to a different model and different evaluation results. Making dataset balanced decreases the total account number as well (2 x number of the organizations). To make a fair comparison, in Humanizr evaluation, they also prepare an unbalanced dataset

Table 4.6: Performance measurements of each model under 7 fold cross-validation.

Model	Dataset	Class	Recall	Accuracy
[8]	Humanizr(Full)	Organization	66.0%	**95.5%**
		Individual	**98.6%**	
Proposed Model	Humanizr(Full)	Organization	**71.8%**	95.0%
		Individual	97.6%	
[8]	Humanizr(Unbalanced)	Organization	**58.6%**	**94.1%**
		Individual	**98.2%**	
Proposed Model	Humanizr(Unbalanced)	Organization	57.6%	92.5%
		Individual	96.9%	
[8]	Humanizr(Balanced)	Organization	**89.4%**	**88.8%**
		Individual	88.2%	
Proposed Model	Humanizr(Balanced)	Organization	85.6%	87.6%
		Individual	**89.6%**	
[3]	Demographer(Full)	Organization	-	94.6%
		Individual	-	
Proposed Model	Demographer(Full)	Organization	**90.4%**	**97.4%**
		Individual	**98.5%**	
[3]	Demographer(Balanced)	Organization	-	85.8%
		Individual	-	
Proposed Model	Demographer(Balanced)	Organization	**95.2%**	**95.0%**
		Individual	**94.8%**	

by keeping the organization and individual account ratio the same as the original dataset but keeping the overall account number the same as the number of accounts in the balanced dataset. Hence, we have followed the same steps to generate the same experiment environments by preparing balanced and unbalanced versions of the Humanizr and balanced version of the Demographer dataset.

In study [15], M3 is trained with their crawled dataset. They treat the entire Humanizr and Demographer datasets as a test set and report the performance. The achieved accuracies are 80.7% for organization and 98.6% for individual accounts, where our model's performances on cross-validation are 90.4% for organization and 98.5% for individual.

Table 4.7: Performance details of the proposed model under 7 fold cross-validation.

Model	Dataset	Class	Precision	Recall	F1 Score	Accuracy
Proposed Model	Humanizr(Full)	Organization	77.4%	71.8%	74.5	95.0%
		Individual	96.8%	97.6%	97.2	
	Humanizr(Unbalanced)	Organization	70.6%	57.6%	63.3	92.5%
		Individual	94.8%	96.9%	95.8	
	Humanizr(Balanced)	Organization	89.3%	85.6%	87.3	87.6%
		Individual	86.3%	89.6%	87.9	
Proposed Model	Demographer(Full)	Organization	90.7%	90.4%	90.6	97.4%
		Individual	98.5%	98.5%	98.5	
	Demographer(Unbalanced)	Organization	88.8%	87.8%	88.3	96.3%
		Individual	97.7%	97.9%	97.8	
	Demographer(Balanced)	Organization	94.8%	95.2%	95	95.0%
		Individual	95.2%	94.8%	95	

Table 4.7 provides the proposed model's evaluation results since we provide additional evaluation metrics such as precision and F1-score besides metrics provided in baseline papers. In the table, the proposed model results on all versions of both Humanizr and Demographer datasets are provided for clarity of performance.

Table 4.8: Comparison of the proposed model using numeric, textual, and textual+numeric features on Demographer full dataset under 7 fold cross-validation.

Features	Class	Precision	Recall	F1 Score	Accuracy
Only Numeric	Organization	52.0%	67.1%	58.6	90.0%
	Individual	96.0%	92.7%	94.3	
Only Textual	Organization	87.7%	**91.6%**	89.6	97.3%
	Individual	**98.7%**	98.1%	98.4	
Numeric+Textual	Organization	**90.7%**	90.4%	**90.6**	**97.4%**
	Individual	98.5%	**98.5%**	**98.5**	

We have also done experiments to determine the effects of the numerical and textual features separately. Since the number of organizational accounts are much less than the individual accounts, the desired model should produce not only high accuracy value, but more importantly, it should also have high precision values for both classes. As it can be seen from Table 4.8, using only the numerical features produce 90% accuracy, but with a very low precision on organizational accounts. On the other hand, using only the textual features produces much higher accuracy and quite high precision values for both classes. Moreover, adding the numerical features to the textual features seems to be increasing accuracy slightly. However, it has a significant effect on increasing the precision of the organizational account by 3%. Therefore, from these results, we can easily claim that the combination of the numeric and the textual features gives the best model.

The results show that the Twitter accounts' metadata contains crucial information about the individual vs. corporation account classification. Textual features make the dataset more separable than the numerical features. Our model has higher scores than the baselines in most of the measures. Using embedding vectors of textual features instead of occurrence information of words for classification gives higher performance.

According to our perspective, misclassification of an organization account is a more severe problem, the accuracy of our proposed model 97.4%, which is the best among baselines.

CHAPTER 5

BIG FIVE PERSONALITY ANALYSIS FROM TWITTER ACCOUNTS

Big five personality traits are well known descriptive measurements for the identification of personalities, also known as OCEAN model. It analyzes personality traits in five different dimensions, which are Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

In psychological studies, many problems and concepts are explained and related to personal traits; therefore, surveys and tests are necessary to measure the traits. Big five personality trait tests have few different versions, and many research studies use this theory to explain the problem and propose a solution. To illustrate, [57] investigates the relation of personality in academic achievement. [58] focuses on burnout and engagement in work and their connection with the Big Five Personality Traits.

Raise of social media platforms provided more personal content and widened the research area. People are spending more and more time on these platforms, and they are even filling personality surveys for the sake of curiosity. Online surveys also made it easier to collect data and conduct research. Studies started to relate social media accounts with big five traits; [59] predicts traits from publicly available information in social media accounts, conversely [60] shows how personality influences social media use. Therefore, it can be said that big five personality trait research field applications are shifted into social media platforms.

5.1 Problem Definition

People share their ideas daily on Twitter and provide lots of information about their life, expectancy, habits, routines, likes, dislikes, and behaviors. This information can be fused to predict the personality traits of a person. Therefore, personality prediction and classification on Twitter accounts became a famous problem. Many studies like [29] and [61] provide their data from Twitter and use Big Five Personality traits as the base methodology for representation and measurement of the personality.

In this chapter, we examine the personality prediction of users from their Twitter account information. We focus on individual Twitter accounts and take advantage of graph structures, account metadata, and user tweets. Set of base users have both Big Five Personality survey results and Twitter account usernames as ground truth.

5.2 Method

This section proposes an end-to-end pipeline starting from data collection to preprocessing and generation of classification models.

5.2.1 Data Collection

The base user group is taken from the study [6] which collects Big Five Personality surveys from about 40 volunteers and applies clustering techniques to predict personality traits. They also had the Twitter usernames of the volunteers. We fetched the metadata of Twitter accounts and last 20 tweets of the account, then we extracted graph features by taking the account as a node in the graph.

We have used Twint [34] open source Twitter scraping tool. Eight of the base user accounts could not be fetched because Twitter accounts either are closed or suspended. Therefore, we collected 32 base user account metadata. Table 5.1 shows the fetched metadata attributes for the base users. There are four text and five numeric attributes selected as metadata and non-discreminative features like id, join date, join time are eliminated. Again, the last 20 tweet texts are scraped using Twint.

Table 5.1: The attributes used in metadata fetching phase.

Profile Attributes	Description	Type
name	Name of the user	text
bio	Description of the account	text
location	User-defined location	text
url	A url provided by user	text
tweets	Number of tweet of the account	integer
following	Number of users this account following	integer
followers	Number of users this account has	integer
likes	Number of tweets user has liked	integer
media	Number of media in the tweets	integer

Table 5.2: The account attributes extracted as graph features.

Feature	Description
FollowingOrgCount	Total organization count among following accounts
FollowingIndCount	Total individual count among following accounts
FollowingTweetsMean	Mean of tweets attribute among following accounts
FollowingFollowersMean	Mean of followers attribute among following accounts
FollowingFollowingMean	Mean of following attribute among following accounts
FollowingLikesMean	Mean of likes attribute among following accounts
FollowingMediaMean	Mean of median attribute among following accounts
FollowersOrgCount	Total organization count among followers accounts
FollowersIndCount	Total individual count among followers accounts
FollowersTweetsMean	Mean of tweets attribute among followers accounts
FollowersFollowersMean	Mean of followers attribute among followers accounts
FollowersFollowingMean	Mean of following attribute among followers accounts
FollowersLikesMean	Mean of likes attribute among followers accounts
FollowersMediaMean	Mean of median attribute among followers accounts
TotalNeighbor	Total count of following and followers user sum.

The graph features are extracted for each base user account. First, we fetched the followers list of the account. For each follower account, scraped the metadata information. After collecting metadata of all followers, we calculated the average values for each numeric feature. The same process is applied to the followings of the account. Table 5.2 shows the extracted features with their description.

5.2.2 Preprocessing

Preprocessing phase can be divided into two parts. The first part is related to numeric features. We have 15 numeric features from graph features and five features from metadata information, which sums up 20 features. Normalization is applied to numeric attributes, and all values are scaled between 0-1. After normalization, they are combined into a vector representation.

The second part is about textual features. We have *name, bio, location, url* textual attributes that come from metadata information. Basic data cleaning techniques such as converting numbers into texts, removing punctuation and stopwords, and lowercasing all texts are applied. After cleaning, words are concatenated and converted into language-independent word embeddings using Sentencepiece model. The preprocessing technique for textual metadata attributes is inherited from the previous chapter, and the same steps are applied because the structure of the data is the same.

The main difference from the previous chapter is that base users' last 20 tweet text is included in the model. Since the tweet frequency of users differs, we wanted to use the most up-to-date tweets in the study. It was an intuitive decision to have only the last 20 tweets because they only represent the last few weeks for most users. Special preprocessing is applied to tweets since their context differs from standard text contents. Tweets are more freestyle texts that include emojis, abbreviations, and informal words. Like the metadata attributes, we removed punctuations and stopwords, lowercased whole text, and turned numbers into texts. Additionally, we replaced URLs with <url>, mentions with <user> and hashtags with <hastag> tags. Because these texts may contain compound words and special names, which does not provide an effective contribution to representation. Also, basic emojis are replaced with specific tags. After these conversions, we convert text into word embeddings using BERT

pre-trained model.

5.2.3 Classification Model

Our model has three main components, which are two MLP layers and an LSTM layer. Numeric features are combined and formed a vector, and fed into an MLP layer. Textual features of metadata are converted into word embeddings and fed into the LSTM layer. Lastly, with the help of a BERT pre-trained model, embeddings are extracted from tweets and fed into the MLP layer. Outputs of all layers are concatenated in a dense layer, and a sigmoid function makes label prediction.

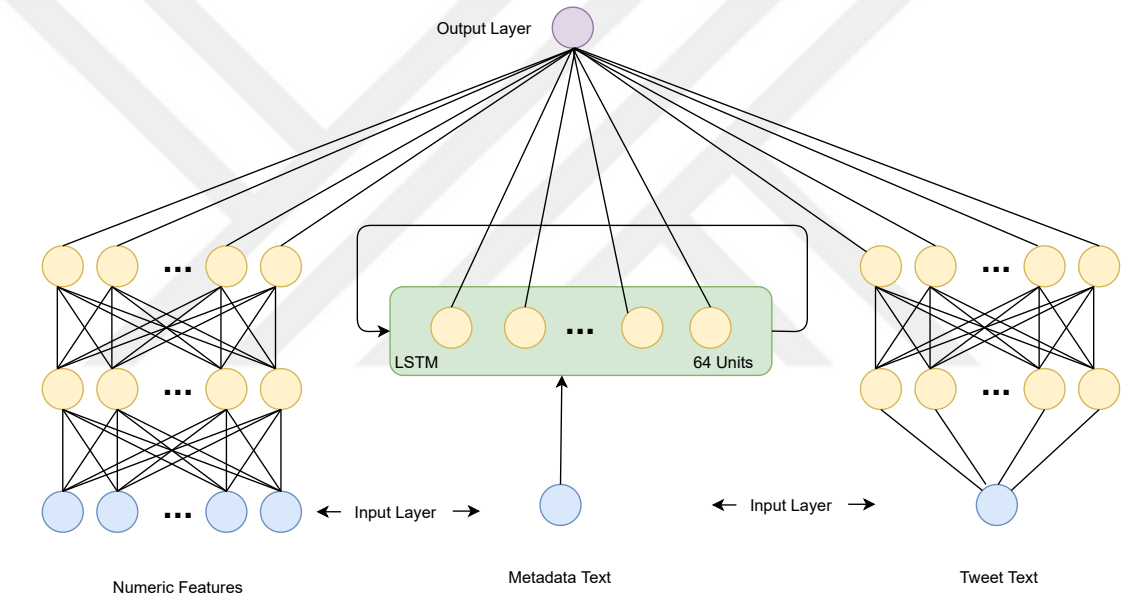


Figure 5.1: Multi-layer perceptron (MLP) for numeric features and long short-term memory (LSTM) for metadata textual features are used. Tweet text is encoded by BERT model and fed into MLP layer. The last layers of the three models are concatenated before connecting it to the output layer. While hidden layers have ReLU as activation function, the output layer uses sigmoid.

We have divided the OCEAN model into five different models. For each dimension, we generate a model. Each model predicts a person is HIGH or LOW in that trait. Hence, the output layer consists of sigmoid functions that can predict the binary label.

5.3 Experiments and Evaluation

We have trained five different models for each OCEAN dimensions. Normally, the Big 5 Personality Test scores each dimension separately between 0-100. However, this score scale is divided into two different areas, which are HIGH and LOW that show personality is dominant or not. HIGH means score is higher than 50 in that dimension, conversely LOW means the score is lower than 50 in the corresponding dimension. Therefore class number for each personality trait is limited by 2.

We have a very limited dataset for model training and testing. As mentioned in the Data Collection section, we have 32 working Twitter account and corresponding Big Five Personality scores. We split 8 of them as a test and trained models with 24 of the accounts.

Table 5.3: Accuracy scores of models corresponding to OCEAN dimensions.

Model	Accuracy
Openness	%62.5
Conscientiousness	%87.5
Extroversion	%75
Agreeableness	%75
Neuroticism	%62.5

Results evaluated due to the data count. Since dataset size is limited, the model cannot learn the features of the OCEAN dimensions exactly. Also, because of the limited training and test sets evaluation metric is kept restricted with accuracy.

CHAPTER 6

CONCLUSION

Social media studies struggle to detect the account type automatically, as it is a prior step in many analysis tasks. Manually processing the dataset is costly and biased so automating the process becomes a necessary study. Previous studies using more detailed information about accounts are language-dependent or need more crawling and training effort. This study aims to classify Twitter accounts as individual vs. organization by eliminating the tweet crawling from the prediction. Using only metadata for the account type classification reduces both the time and the space complexity. Moreover, one can not directly access the protected accounts' detailed information, but these accounts' metadata are publicly available. The experiments reveal that it is possible to classify account type using only the metadata, and the proposed model gets ahead of the baselines. As future work, one possible extension is to apply the feature extraction to the numerical features instead of using them directly. Our proposed method gives considerably high-performance scores as a multi-lingual model. There is a chance to boost the performance using accounts with the same language in the study. Besides, in addition to the tool that we have provided that can be used to determine whether a given tweet account is individual or organization, we also plan to provide our proposed method as a service to make it a quick but efficient preprocessing tool for researchers working on Twitter.

Personality prediction is a popular topic in the social media domain, and it requires voluntary personal data. Since our personal survey count is very limited, we could not reach the results we desired. However, with the small dataset we have, we followed a separation strategy and divided the problem into five different classification problems. For each OCEAN dimension, we trained a single model, which provided

more reliable and promising results. As future work, we are planning to find more volunteers for the Big Five Personality survey and want to train our proposed model with a higher amount of data. Another future plan is to train a single model for all dimensions and provide a more comprehensive solution for the personality classification problem.

REFERENCES

[1] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, "Our twitter profiles, our selves: Predicting personality with twitter," in *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pp. 180–185, IEEE, 2011.

[2] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[3] Z. Wood-Doughty, P. Mahajan, and M. Dredze, "Johns hopkins or johnny-hopkins: Classifying individuals versus organizations on twitter," in *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pp. 56–61, 2018.

[4] J. Pennington, "Global vectors for word representation." `https://nlp.stanford.edu/projects/glove`. Accessed: 2020-10-19.

[5] "Twitter by the numbers: Stats, demographics fun facts." `https://www.omnicoreagency.com/twitter-statistics`. Accessed: 2020-106-17.

[6] E. Tutaysalgir, P. Karagoz, and I. H. Toroslu, "Clustering based personality prediction on turkish tweets," in *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 825–828, IEEE, 2019.

[7] L. De Silva and E. Riloff, "User type classification of tweets with implications for event recognition," in *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pp. 98–108, 2014.

[8] J. McCorriston, D. Jurgens, and D. Ruths, "Organizations are users too: Char-

acterizing and detecting the presence of organizations on twitter," in *Ninth International AAAI Conference on Web and Social Media*, 2015.

[9] I. Samborskii, A. Filchenkov, G. Korneev, and A. Farseev, "Person, organization, or personage: Towards user account type prediction in microblogs," in *International Conference on Electronic Governance and Open Society: Challenges in Eurasia*, pp. 111–122, Springer, 2018.

[10] K. E. Daouadi, R. Z. Rebaï, and I. Amous, "Organization vs. individual: Twitter user classification," in *Proceedings of the second Conference on Language Processing and Knowledge Management*, vol. 2279, 2018.

[11] G. Tavares and A. Faisal, "Scaling-laws of human broadcast communication enable distinction between human, corporate and robot twitter users," *PloS one*, vol. 8, no. 7, 2013.

[12] G. Tavares, S. Mastelini, *et al.*, "User classification on online social networks by post frequency," in *Anais do XIII Simpósio Brasileiro de Sistemas de Informação*, pp. 464–471, SBC, 2017.

[13] S. M. Kim, C. Paris, R. Power, and S. Wan, "Distinguishing individuals from organisations on twitter," in *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 805–806, 2017.

[14] S. Alzahrani, C. Gore, A. Salehi, and H. Davulcu, "Finding organizational accounts based on structural and behavioral factors on twitter," in *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pp. 164–175, Springer, 2018.

[15] Z. Wang, S. A. Hale, D. I. Adelani, P. A. Grabowicz, T. Hartmann, F. Flöck, and D. Jurgens, "Demographic inference and representative population estimates from multilingual social media data," in *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019* (L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, and L. Zia, eds.), pp. 2056–2067, ACM, 2019.

[16] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2261–2269, IEEE Computer Society, 2017.

[18] N. A. AlSomaikhi and Z. A. Alzamil, “Twitter users’ classification based on interest: Case study on arabic tweets,” *International Journal of Information Retrieval Research (IJIRR)*, vol. 10, no. 1, pp. 1–12, 2020.

[19] D. Karatay and P. Karagoz, “User interest modeling in twitter with named entity recognition,” in *5th Workshop on Making Sense of Microposts*, 2015.

[20] M. A. Raghuram, K. Akshay, and K. Chandrasekaran, “Efficient user profiling in twitter social network using traditional classifiers,” in *Intelligent systems, technologies and applications*, pp. 399–411, Springer, 2016.

[21] A. Souri, S. Hosseinpour, and A. M. Rahmani, “Personality classification based on profiles of social networks’ users and the five-factor model of personality,” *Human-centric Computing and Information Sciences*, vol. 8, no. 1, p. 24, 2018.

[22] T.-Y. Chen, Y.-M. Chen, and M.-C. Tsai, “A status property classifier of social media user’s personality for customer-oriented intelligent marketing systems: Intelligent-based marketing activities,” *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 16, no. 1, pp. 25–46, 2020.

[23] E. Colleoni, A. Rozza, and A. Arvidsson, “Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data,” *Journal of communication*, vol. 64, no. 2, pp. 317–332, 2014.

[24] Y. M. Çetinkaya, İ. H. Toroslu, and H. Davulcu, “Developing a twitter bot that can join a discussion using state-of-the-art architectures,” *Social Network Analysis and Mining*, vol. 10, no. 1, pp. 1–21, 2020.

[25] R. Napoli, A. M. Ertugrul, A. Bozzon, and M. Brambilla, “A user modeling pipeline for studying polarized political events in social media,” in *International Conference on Web Engineering*, pp. 101–114, Springer, 2018.

[26] J. S. Alowibdi, U. A. Buy, and P. Yu, "Language independent gender classification on twitter," in *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*, pp. 739–743, 2013.

[27] M. Vicente, J. P. Carvalho, and F. Batista, "Using unstructured profile information for gender classification of portuguese and english twitter users," in *International Symposium on Languages, Applications and Technologies*, pp. 57–64, Springer, 2015.

[28] L. Liu, D. Preotiuc-Pietro, Z. R. Samani, M. E. Moghaddam, and L. Ungar, "Analyzing personality through social media profile picture choice," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 10, 2016.

[29] M. Skowron, M. Tkalčič, B. Ferwerda, and M. Schedl, "Fusing social media cues: personality prediction from twitter and instagram," in *Proceedings of the 25th international conference companion on world wide web*, pp. 107–108, 2016.

[30] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the national academy of sciences*, vol. 110, no. 15, pp. 5802–5805, 2013.

[31] C. Donnat, M. Zitnik, D. Hallac, and J. Leskovec, "Learning structural node embeddings via diffusion wavelets," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1320–1329, 2018.

[32] L. F. Ribeiro, P. H. Saverese, and D. R. Figueiredo, "struc2vec: Learning node representations from structural identity," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 385–394, 2017.

[33] K. Cherry, "What are the big 5 personality traits?." `https://www.verywellmind.com/the-big-five-personality-dimensions-2795422`. Accessed: 2020-07-14.

[34] F. Poldi, "Twint-twitter intelligence tool." https://github.com/twintproject/twint. Accessed: 2020-05-21.

[35] J. Brownlee, "What are word embeddings for text?." https://machinelearningmastery.com/what-are-word-embeddings/. Accessed: 2020-09-07.

[36] B. Heinzerling and M. Strube, "BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), 2018.

[37] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, 2016.

[38] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Association for Computational Linguistics, 2018.

[39] P. Gage, "A new algorithm for data compression," *C Users Journal*, vol. 12, no. 2, pp. 23–38, 1994.

[40] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[41] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Association for Computational Linguistics, 2019.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017.

[43] M. Mohd, R. Jan, and N. Hakak, "Enhanced bootstrapping algorithm for automatic annotation of tweets," *Int. J. Cogn. Informatics Nat. Intell.*, vol. 14, no. 2, pp. 35–60, 2020.

[44] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?," *IEEE Trans. Dependable Secur. Comput.*, vol. 9, no. 6, pp. 811–824, 2012.

[45] L. S. Martinez, M.-H. Tsou, and B. H. Spitzberg, "A case study in belief surveillance, sentiment analysis, and identification of informational targets for e-cigarettes interventions," in *Proceedings of the 10th International Conference on Social Media and Society*, pp. 15–23, 2019.

[46] C. Lu, W. Lam, and Y. Zhang, "Twitter user modeling and tweets recommendation based on wikipedia concept graph," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[47] A. García-Silva, V. Rodríguez-Doncel, and Ó. Corcho, "Semantic characterization of tweets using topic models: A use case in the entertainment domain," *Int. J. Semantic Web Inf. Syst.*, vol. 9, no. 3, pp. 1–13, 2013.

[48] S. K. Bharti, R. Pradhan, K. S. Babu, and S. K. Jena, "Sarcastic sentiment detection based on types of sarcasm occurring in twitter data," *Int. J. Semantic Web Inf. Syst.*, vol. 13, no. 4, pp. 89–108, 2017.

[49] N. V. Patel and H. Chhinkaniwala, "Investigating machine learning techniques for user sentiment analysis," *Int. J. Decis. Support Syst. Technol.*, vol. 11, no. 3, pp. 1–12, 2019.

[50] J. R. Alharbi and W. S. Alhalabi, "Hybrid approach for sentiment analysis of twitter posts using a dictionary-based approach and fuzzy logic methods: Study

case on cloud service providers," *Int. J. Semantic Web Inf. Syst.*, vol. 16, no. 1, pp. 116–145, 2020.

[51] F. Atefeh and W. Khreich, "A survey of techniques for event detection in twitter," *Computational Intelligence*, vol. 31, no. 1, pp. 132–164, 2015.

[52] D. Ramachandran and R. Parvathi, "Enhanced event detection in twitter through feature analysis," *International Journal of Information Technology and Web Engineering (IJITWE)*, vol. 14, no. 3, pp. 1–15, 2019.

[53] R. Srivastava and M. P. S. Bhatia, "Real-time unspecified major sub-events detection in the twitter data stream that cause the change in the sentiment score of the targeted event," *International Journal of Information Technology and Web Engineering (IJITWE)*, vol. 12, no. 4, pp. 1–21, 2017.

[54] C. Toraman, "Early prediction of public reactions to news events using microblogs," in *Seventh BCS-IRSG Symposium on Future Directions in Information Access 7*, pp. 1–4, 2017.

[55] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *arXiv preprint arXiv:1310.4546*, 2013.

[56] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[57] M. Komarraju, S. J. Karau, R. R. Schmeck, and A. Avdic, "The big five personality traits, learning styles, and academic achievement," *Personality and individual differences*, vol. 51, no. 4, pp. 472–477, 2011.

[58] H. J. Kim, K. H. Shin, and N. Swanger, "Burnout and engagement: A compar-

ative analysis using the big five personality dimensions," *International Journal of Hospitality Management*, vol. 28, no. 1, pp. 96–104, 2009.

[59] J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media," in *CHI'11 extended abstracts on human factors in computing systems*, pp. 253–262, 2011.

[60] G. Seidman, "Self-presentation and belonging on facebook: How personality influences social media use and motivations," *Personality and individual differences*, vol. 54, no. 3, pp. 402–407, 2013.

[61] V. Ong, A. D. Rahmanto, D. Suhartono, A. E. Nugroho, E. W. Andangsari, M. N. Suprayogi, *et al.*, "Personality prediction based on twitter information in bahasa indonesia," in *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pp. 367–372, IEEE, 2017.