IDENTIFYING PERSONALITY AND TOPICS OF SOCIAL MEDIA

A THESIS IN
Computer Science

Presented to the Faculty of the University
of Missouri-Kansas City in partial fulfillment
of the requirements for the degree

MASTER OF SCIENCE

By
TRINADHA R MUPPALA

MCA, Motilal Nehru National Institute of Technology Allahabad, India

Kansas City, Missouri
2019

IDENTIFYING PERSONALITY AND TOPICS OF SOCIAL MEDIA

Trinadha R Muppala, Candidate for the Master of Science Degree

University of Missouri-Kansas City, 2019

ABSTRACT

Twitter and Facebook are the renowned social networking platforms where users post, share, interact and express to the world, their interests, personality, and behavioral information. User-created content on social media can be a source of truth, which is suitable to be consumed for the personality identification of social media users. Personality assessment using the Big 5 personality factor model benefits organizations in identifying potential professionals, future leaders, best-fit candidates for the role, and build effective teams. Also, the Big 5 personality factors help to understand depression symptoms among aged people in primary care. We had hypothesized that understanding the user personality of the social network would have significant benefits for topic modeling of different areas like news, towards understanding community interests, and topics.

In this thesis, we will present a multi-label personality classification of the social media data and topic feature classification model based on the Big 5 model. We have built the Big 5 personality classification model using a Twitter dataset that has defined openness, conscientiousness, extraversion, agreeableness, and neuroticism. In this thesis, we (1) conduct personality detection using the Big 5 model, (2) extract the topics from Facebook and Twitter data based on each personality, (3) analyze the top essential topics, and (4) find the relation between topics and personalities. The personality would be useful to identify

what kind of personality, which topics usually talk about in social media. Multi-label classification is done using Multinomial Naïve Bayes, Logistic Regression, Linear SVC. Topic Modeling is done based on LDA and KATE. Experimental results with Twitter and Facebook data demonstrate that the proposed model has achieved promising results.

APPROVAL PAGE

The faculty listed below, appointed by the Dean of the School of Computing and Engineering, have examined a thesis titled "Identifying Personality and Topics of Social Media" presented by Trinadha R Muppala, candidate for the Master of Science degree, and hereby certify that in their opinion, it is worthy of acceptance.

Supervisory Committee

Yugyung Lee, Ph.D. (Committee Chair)
Department of Computer Science Electrical Engineering

Praveen Rao, Ph.D.
Department of Computer Science Electrical Engineering

Ye Wang, Ph.D.
Department of Communication Studies

Contents

LIST OF ILLUSTRATIONS

Figure Page

LIST OF TABLES

Table Page

ACKNOWLEDGMENTS

I would like to thank Dr. Yugyung Lee for her valuable guidance and immense support throughout the research work as my advisor. Data analytics growing very fast Dr. Lee always keeps herself up to date with the latest research and encourages her students to work towards cut edge technologies. I am amazed by her positive energy and patience.

I would like to thank UMKC professors, Dr. Praveen Rao, Dr. Deepak Medhi and Dr. Sing Song for their extraordinary teaching skills making all the courses learn easy and making students successful in the Data Science stream. All the knowledge I have gained at UMKC will help my carrier growth; I am very happy picking UMKC to do my Masters.

I would like to thank Mayanka Chandrashekar, Ph.D. student for teaching Data analytics and Knowledge discovery management. She helped in changing my way of thinking from software engineering to analytics that helped me a lot with my thesis and I would like to thank Saria Goudarzvand, Ph.D. student for her guidance for my thesis. Her skills in machine learning and deep learning are amazing.

Finally, I would like to thank my husband Ramarao Chavali, my kids Vamsi Chavali, Rima Chavali, and Rahul Chavali for their support without their help and support I will not be able to finish my thesis. Thanks to my father Sarma Muppala and my friends for their encouragement and to my colleagues for covering when I am at school during business hours.

CHAPTER 1. INTRODUCTION

People's patterns of thinking, feelings, behavior reflect their personality. Psychological researches have been performed to identify different personality identification models. Combining those results with Artificial intelligence will benefit from identifying human thinking. Natural language processing (NLP) allows the machine to interpret natural language by utilizing the power of artificial intelligence, computational linguistics, and computer science.

Text classification is becoming increasingly important as it allows to get insights. Topic detection is one of the important automatic text classifications. Topic modeling discovers the hidden topical patterns in the collection of textual information. Topic modeling is unsupervised learning that builds clusters of the words. There are several approaches available for finding out topics from text corpus namely Latent Dirichlet Allocation (LDA) [1], K-Competitive autoencoder (KATE) [2].

There are different personality models MBTI (Myers-Briggs personality types), DISC (Dominance Influence Steadiness Conscientiousness), Big 5 model. The Big 5 [3] model has proven as the best model for psychometric testing. Big 5 model also referred to OCEAN, personality traits for the Big 5 model are Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism. OCEAN is reliable, valid, applied widely in the professional world, and universally accepted by researchers.

	Dream5 (OCEAN)	DISC	MBTI
Measured Traits	**30 (spectrum)**	4 (quadrant)	4 (binary)
Results	∞	12 profiles	16 personalities
Valid?	**Yes** *FBI Correlation (Jonson, 2014)*	Yes	No *Validity non measurable (Stein & Swan, 2019)*
Reliable?	**Yes** *IPIP NEO (Jonson, 2014)*	Yes	No *Self perception only (Stein & Swan, 2019)*
Predictive?	**Yes** *Predictor in performance, satisfaction, training, etc. (Juhász, 2010; Locander, Mulki, & Weinberg, 2014, Ziegler, Bensch, Maaß, Schult, Vogel, & Bühner, 2014; Kiang, 2012; Judge et al, 2013)*	No *Not recommended for prediction (discprofile.com, n.d.) Measures behavior, not personality (Wolfe, 2011) Inadequate (Envisia Learning, 2018)*	No *Inadequate for individual prediction (Stein & Swan, 2019)*

Figure 1: Personality Models Comparison for Recruitment [21]

Personality assessment using the Big 5 (OCEAN) factor model benefits organizations in identifying potential professionals, future leaders, best fit candidates for the role and build effective teams. The Big 5 personality factors help to understand depression symptoms [4] among aged people in primary care. Studies have proven that personality with high conscientiousness was associated with likely identification of depression by primary care physicians. Figure 1 shows the OCEAN is a more reliable and predictive personality assessment model for the recruitment.

Different characteristics of each personality with high and low traits defined as following. People with Openness personality with the high trait are imaginative, preference for variety, Independent, happy to think about abstract concepts, and with lot traits are practical, preference for routine, dislikes abstract or theoretical, confirming. People with Conscientiousness personality with the high trait are organized, careful, disciplined and with

low trait dislikes structure and schedules, disorganized, careless, impulsive fails to complete necessary or assigned tasks. People who are Extraversion with high trait enjoys being the center of attention, fun-loving, affectionate and with low trait feel exhausted when having to socialize a lot, retiring, sober, reserved. People with Agreeableness personality with the high trait have a great deal of interest in other people, soft-hearted, trusting, assists others who need help and lot trait takes little interest in others, ruthless, suspicious, uncooperative. People with Neuroticism personality with high trait experiences a lot of stress, worries about many different things, gets upset easily, experiences dramatic shifts in mood, feel anxious, struggles to bounce back after stressful events and with low trait emotionally stable, deal well with stress, rarely feels sad or depressed, doesn't worry much, Is very relaxed.

Twitter and Facebook are the renowned social networking platforms where users post, share, interact and express to the world, their details, interests, personality and behavioral information. User-created content on social media can be a rich source of data that can be utilized for the identification of the personality of the author. In this thesis, we present a multi-label classification of tweets/Facebook status. Also, extract the topics from the Twitter data based on each personality find the relation between the topic's users discussed and their personality.

1.1 Problem Statement

Twitter identified as the most used social platform in 2013, in 2018 Twitter got a monthly active user count of 321 million [5]. Twitter can be viewed as "SMS of the Internet". Users post, interact, express and share their personality and behavioral information on Twitter. Tweets are unsupervised data; Twitter users discuss different topics in social media. There are

different researches have been done on personality assessment, models like BIG 5, DISC provide personality assessment. The five personality traits in the Big 5 Model [3] are Openness to Experience, Conscientiousness, Extraversion, Agreeableness, Neuroticism. For this thesis, we will work on identifying Twitter user personality and topics Twitter users are talking about and find the connection or relation between topics usually people discuss about and the personality to conclude which kind of personality which topics usually interested about in social media. A personality trait is a multi-label classification problem, existing papers provided the separate models for each personality.

1.2 Proposed Solution

The proposed solution focuses on the classification of the unsupervised Twitter data and identifies the topics of Twitter users discussing and relate the topics with personality. In this thesis we present

1. Build Multi-label Big 5 [3] personality classification using Facebook, Twitter data
2. Identify topics for each personality using KATE [2]
3. Build Multi-label Classification Model based on LDA [1] Topic Features
4. Find the relation between topics and personalities

David Stillwell and Michal Kosinski did project myPersonality [6] created a Facebook app for users to participate in psychological research, myPersonality is multi labeled data each status text is labeled with y/n for each personality. We used myPersonality data for personality classification modeling and used that model to derive personality for Twitter

users. We used KATE [2] for topic modeling extracted the topics for each personality of the tweets. Extracted the features of each text and built a model for personality classification.

CHAPTER 2. BACKGROUND AND RELATED WORK

This chapter gives the background information of various components used in the thesis and gives an overview of related work that will help in understanding this work better.

2.1 Related Work

Different research has been done on personality computing using different models like OCEAN, DISC. Personality classification multi-label classification. Topic modeling has been done using Twitter data to provide insights into different areas like football news, natural disaster.

2.1.1 Personality Identification

Research has been done by computing Personality traits; user personality data collected through personality-related survey [7] created a Twitter application with 45-question version of the Big Five personality inventory. Processed the date through Linguistic Inquiry and Word Count (LIWC) which produced 79 text features, ran the text through MRC Psycholinguistic Database. Performed word by word sentiment analysis that assigns words sentiment values -1 to +1 scale. Two regression algorithms Gaussian Process and Zero R used each with 10-fold cross-validation with 10 iterations. They have identified neuroticism most difficult and Openness easiest to compute. Predict score within 11% - 18% predict the score on each of Five personality Traits.

Research has been done computing Personality traits from Tweets [8]. Used Facebook project data myPersonality user status trained five different Support Vector Machine (SVM) each for one personality trait. They defined hyperparameters (Kernel, C, Gamma, Degree) to optimize the model by minimizing the loss function (MSE). Trained the model using myPersonality data and extended it to identify personality traits from the tweets. Twitter users answered the psychological questionnaire.

Developed a document modeling technique based on CNN features [9] extractor for personality classification on essay data. Essays data consists of a total of 2,468 essays or daily writing submissions from 34 psychology students. Fed sentences from the essays to convolution filters to obtain the sentence model in the form of n-gram feature vectors. They represented each essay by aggregating the vectors of its sentences. They concatenated the obtained vectors with the Mairesse features which were extracted from the texts directly at the preprocessing stage; which improved the method's performance. Discarding emotionally neutral input sentences from the essays further improved the results. For final classification, they fed the document vector into a fully connected neural network with one hidden layer. Their results outperformed the current state of the art for all five traits.

Research on the Retaliation ship between Emoji usage patterns and Big Five personality traits done [10] by using 352,245 Twitter users collected from March 2016 to June 2016, kept users whose number of followers less than 1649 and the number of followees less than 1180. Each user's tweets submitted to the Receptiviti LIWC model to obtain a user's psychological attributes. Found high correlation values and specific emoji usage in line with perceived user's personality traits. Found Openness shows no relationship with emoji usage.

Personality identification was done based on the DISC model. DISC performs behavioral assessment keeping four principal and key behaviors which are: Dominance, Influence, Stability, and Compatibility. Research has been done to identify DISC personality characteristics [11]. One million tweets downloaded using keywords related to DISC four behaviors. The text mining technique (clustering) and sentiment analysis are performed and mapped results to DISC personality characteristics.

2.1.2 Topics Identification from Tweets

Topic Modeling performed on the aggregated tweets conversation [12]. For the data a document consists of a seed tweet, all the tweets written in reply to it posted by other users and the replies of the original poster to these. Gathered data for 14 topics (Food, Science, Books, Business. Technology, Art, Health, Fashion. Charity. Entertainment, Politics, News, Music, Sports), for each topic select 25 most influential users. Data trained on LDA and ATM topic models. ATM model outperforms LDA.

Using the Twitter data of football, news provided on different topics people discussed on Twitter [13]. Data retrieved from reliable Indonesian Twitter accounts that posted about football news. These accounts include @bolanet, @detiksport, @goal_id, @panditfootbal, @vivabola. The total data obtained from those Twitter accounts are 120,639 tweets with a period from 1st January 2017 to 24th December 2017. Using Latent Dirichlet Allocation method – LDA obtained several insightful topics such as pre-match analysis, live match update, football club achievements. For this thesis using LDA[1], KATE[2] topic models provided the different topics discussed by the social media users for each personality.

2.1.3 Text Classification using Machine learning

Automatic text classification have been extensively studied and rapid progress seems in this area, including the machine learning approaches [14] such as Bayesian classifier, Decision Tree, K-nearest neighbor(KNN), Support Vector Machines(SVMs), Neural Networks, Latent Semantic Analysis, Rocchio's Algorithm, Fuzzy Correlation and Genetic Algorithms. Supervised learning techniques are used for automatic text classification, where pre-defined category labels are assigned to documents based on the likelihood suggested by a training set of labelled documents.

Rocchio's Algorithm is a vector space method for document routing or filtering in informational retrieval, build prototype vector for each class using a training set of documents. The k-nearest neighbor algorithm (k-NN) is used to test the degree of similarity between documents and k training data and to store a certain amount of classification data, thereby determining the category of test documents.

Decision rules classification method uses the rule-based inference to classify documents to their annotated categories. Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' Theorem with strong independence assumptions. The SVM to seek for the decision surface that best separates the positive from the negative data in the n-dimensional space, so called the hyper plane. The document representatives which are closest to the decision surface are called the support vector. For this thesis multi-label personality classification done using Multinomial Naïve Bayes, Logistic Regression and Linear SVC.

CHAPTER 3. PROPOSED FRAMEWORK

Personality identification of social media text is multi-label classification Multi-label classification done by wrapping Multinomial Naïve Bayes [15], Logistic Regression [16], Linear SVC [17] in OneVsRestClassifier [18]. Build topic featured based multi-label classification using LDA. LDA [1] provides the features of the text which can be utilized for the classification model. KATE - K-Competitive Autoencoder [2] to extract topics from Twitter/Facebook data. KATE model provides that it can learn Semantically meaningful representations from the text, used KATE to extract topics from Facebook and Twitter data for each personality. Experimental results with Twitter and Facebook data demonstrate that the proposed model has achieved promising results.

3.1 Framework Architecture

The architecture diagram shown in Figure 2 portrays the full architecture multi-label personality classification model and topic modeling. Facebook myPersonality data used to build a multi-label personality classification. The Facebook dataset is multi-labeled with each status labeled with one or more personality. Facebook has been collected for research purposes by David Stillwell and Michael Kosinski [6] a Facebook application that administered a personality test and collected a wide range of personal and activity information from Facebook's profile of users mainly from the US and UK under their consent. Collected Twitter data for random 28 users and President Trump tweets for the year of 2018.

Text data should be cleaned through Natural language processes and vectorized to build the classification model. TFIDF, Word2Vec used to vectorize the data, extracted topic

features from LDA [1] a topic model to build topic-based classification model. KATE topic modeling applied to identify the topics discussed for each personality.

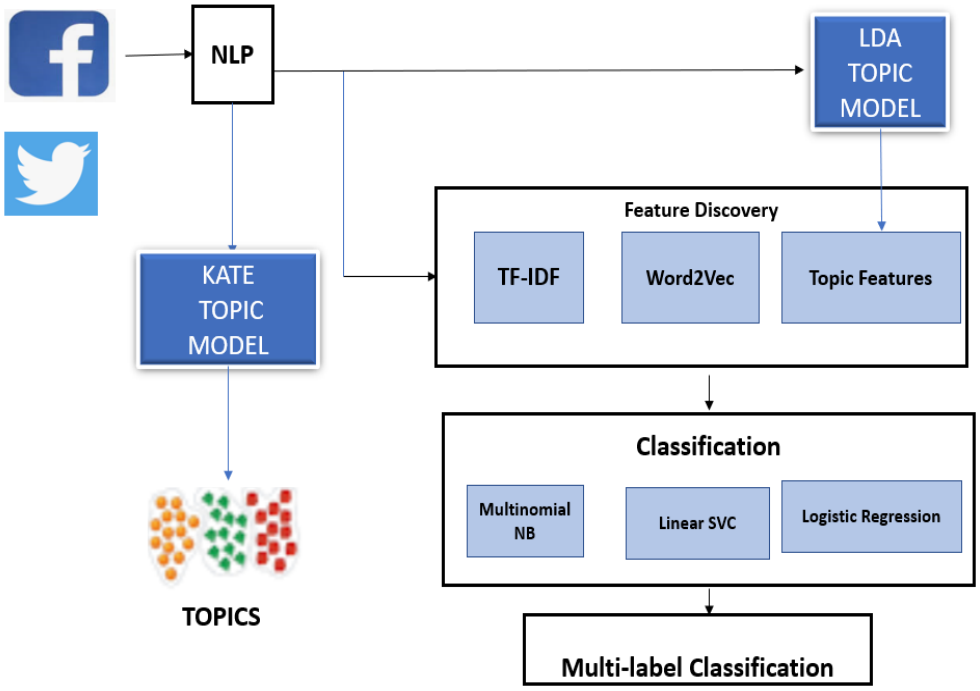

Figure 2: Architecture of Classification and Topic Modeling

Data cleaned using the natural language processing, removed stop words and punctuation, URLs, hashtags, UTF characters from the text. Converted all the text to lowercase. Table 1 provides the data before cleaning and after cleaning the text data

Table 1: Data Before and After Cleaning

Original text	Cleaned text
Dukes are looking to defend their title while Bison are looking to reclaim it NDSU last won in were eliminated in Ã¢Â€Â˜ by JMU	dukes defend title bison reclaim ndsu eliminated jmu
WeÃ¢Â€Â™re here and hyped for the #FCSChampionship No NDSUfootball takes on No JMUFootball in Frisco at CT	hyped ndsufootball takes jmufootball frisco
Jeremiah Briscoe wins his nd straight Walter Payton Award FCS Heisman Only the second player in FCS history to win after Armanti Edwards	jeremiah briscoe wins straight walter payton award fcs heisman player fcs history win armanti Edwards
QB Jarrett Stidham will forgo NFL draft return to Auburn	jarrett stidham forgo nfl draft auburn
WVU freshman safety Derrek Pitts cited for carrying a concealed weapon underage outside HS basketball game	wvu freshman safety derrek pitts cited carrying concealed weapon underage basketball game
The Bills Postseason TD Drought turns tomorrow Expected to declare next month on National Signing Day #BUFvsJAX	bills postseason drought turns tomorrow expected declare month national signing day

Finding Word frequencies by far the most popular method is called TFIDF. This is an acronym that stands for "Term Frequency – Inverse Document" Frequency which are the components of the resulting scores assigned to each word. Term Frequency: This summarizes how often a given word appears within a document. Inverse Document Frequency: This downscales words that appear a lot across documents. Converted the Facebook/Twitter data to TFIDFVectorizer which is a matrix of TFIDF Features.

Word2Vec is a neural net which is two-layer process text data. Word2Vec take text as input and provides numeric vectors set as output. Word2Vec got 2 models

i. Continuous bag of words and

ii. Skip-gram model.

A continuous bag of words model can be used to predict the word in the sentence by the context. Skip-gram model predicts the context of the word in a sentence.

Table 2: Text with Multi-label Personality

STATUS	cOPN	cCON	cEXT	cAGR	cNEU
Um, amy poehler is rocking my world right now. seriously, the bad date rant was HYSTERICAL. tivo that shit. and by that shit, i mean parks and recreation.	0	1	1	0	1
Ten Movies to Watch Right Now (and some you can Instant Netflix) 1. La Vie En Rose 2. Shrink (if you love LA) 3. Paris Je'taime (if you love Paris) 3. Clay Pidgeons (*PROPNAME* is priceless) 4. Quills 5. Away We Go 6. Sunshine Cleaning 7. A League of Their Own 8. Smart People (I Heart *PROPNAME*'s Page and SJP) 9. Frost//Nixon 10. Doubt	0	1	1	0	1
has GOT to stop waking up at 1pm...	1	0	1	1	0
is not feeling exactly top-notch...	1	0	1	1	0
is sore and wants the knot of muscles at the base of her neck to stop hurting. On the other hand, YAY I'M IN ILLINOIS! <3	1	0	0	0	1

Using different Machine learning algorithms-built separate binary classification models for each personality. Multinomial Naïve Bayes [15], Logistic Regression [16], Linear SVC [17] wrapped with One vs Rest Classifier and compared the mean squared error of the models.

But personality classification is not a binary classification. Each text is labeled with multiple personalities as shown in Table 2. For multi-label classification labels should MultiLabelBinarizer the personality labels to 1 or 0. Facebook data is multi-label data trained Facebook data and build multi-label the classification model and derived topics for each personality Figure 3 provides the architecture of Facebook classification and topic model. Data cleaned using the NLP and text data vectorized using TFIDF/Wrod2Vec. Multilabel

classification using Multinomial Naïve Bayes, Logistic Regression, Linear SVC. Separate topic models built for each personality.

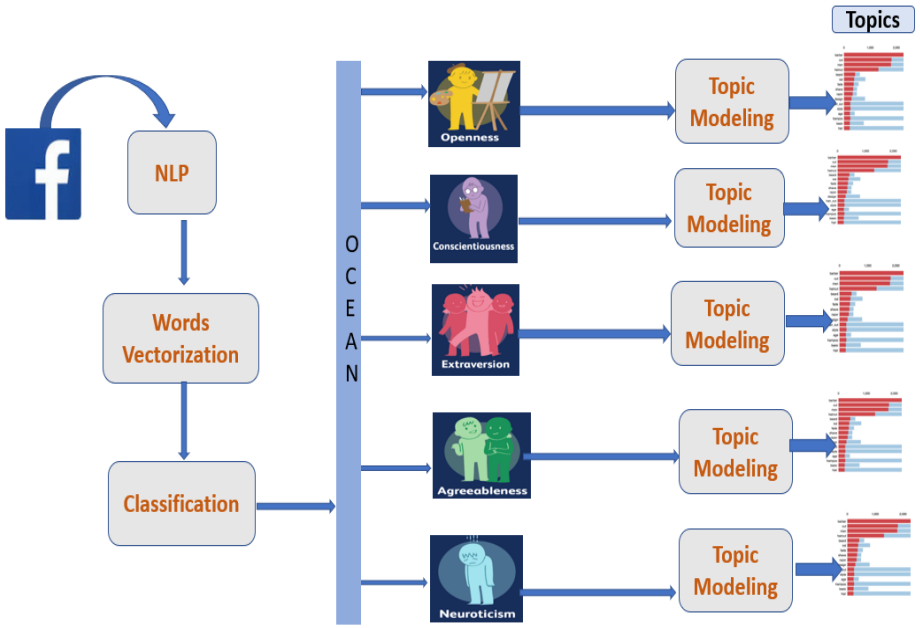

Figure 3: Facebook Multi-label Classification Topic Modeling

Multinomial Naïve Bayes [15] is conditional independence where feature probabilities are independent given the class *c*. Multinomial Naive Bayes [15] equation is the parametric model used for text classification, wherein a document 'd' word $'w_i'$ occurs in $'f_i'$ times.

$$P\left(\frac{c}{d}\right) = \frac{P(c)\prod_{i=1}^{n} P(W_i/c)^{f_i}}{P(d)}$$

For a given class value c the condition probability P(w_i/c) is how many times the word happens to be there in a document 'd' where 'n' represents the number of total unique words

in document ‘d’. The prior probability P(c) that a document happens for the class label ‘c’ in the collection of documents

Logistic Regression [16] analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name logistic regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. In logistic regression, a mathematical model of a set of explanatory variables is used to predict a *logit* transformation of the dependent variable.

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

If p is the proportion of observations with an outcome of 1, then 1-p is the probability of an outcome of 0. The ratio p/(1-p) is called the odds and the logit is the logarithm of the odds or just log odds.

A Support Vector Machine (SVM) performs classification by finding the hyperplane that maximizes the margin between the two classes. The vectors that define the hyperplane are the support vectors. Linear SVC kernel is a Linear function that finds the hyperplane that maximizes the margin and minimizes the misclassifications. Linear Kernel is the product of two vectors $\vec{x}_i$, $\vec{x}_j$

$$K\left(\vec{x}_{i,}\vec{x}_j\right) = \vec{x}_i \,.\vec{x}_j$$

3.2 Topic Discovery

In this thesis, we conducted the topic modeling to extract the hidden topics in the short text for each personality. Here we used KATE [2] which is a generative model for topic discovery. K-Competitive Autoencoder for Text is an unsupervised, statistical method to document modeling that learns latent semantic topics in collections of text documents. KATE introduces competition among neurons. Input and hidden neurons and hidden and output neurons are fully connected. LDA [1] points out that words for each personality and documents discussing similar topics will use a similar group of words. Latent topics are thus revealed by finding groups of words in the corpus that commonly occur together within documents. LDA provides topic feature extraction which is used for topic feature-based multi-label classification.

Twitter data labeled with the Facebook classification model defined in Figure 4, trained with different algorithms Multinomial Naïve Bayes [15], Logistic Regression [16], Linear SVC [17] and compared the models. Using KATE [2] derived topics for each functionality identified the relationship between the topics and the personality traits.

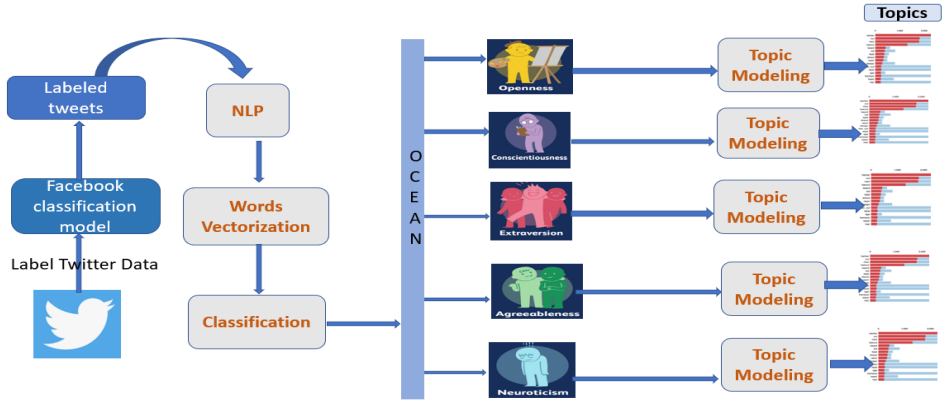

Figure 4: Twitter Multi-label Classification Topic Modeling

CHAPTER 4. RESULTS AND EVALUATIONS

4.1 Introduction

Facebook, Twitter data converted to numeric vectors by different approaches TFIDFVectorizer which is a matrix of TF-IDF Features and Word2Vec vector. Using the Pipeline approach wrapped machine learning algorithms Multinomial Naïve Bayes [15], Logistic Regression [16], Linear SVC [17] into OneVsRestClassifier [18] for multi-label personality classification. Evaluated the results with different measurements like precision, recall, F1-score [19]. Using KATE [2] derived the topics for each personality for Facebook, Twitter data.

4.2 Data Preparation

Facebook [6] and Twitter social media data used for personality classification using the Big 5 model [3]. Data is cleaned and vectorized before processing through machine and topic learning algorithms.

4.2.1 Twitter Data Collection

The tweets are collected for random 7 days for the year 2018. From those tweets picked random 29 users including President Trump, collected all the tweets for those users for the year 2018, Cleaned data using Natural Language Process removed stop words, URLs.

4.2.2 Facebook and Twitter Data details

Table 3 provides details of the Facebook data [6]. Facebook status collected from total number 250 users, provided Facebook status count before cleaning the data, and after

cleaning the data using Natural language techniques. Statuses removed if the number of words less than 4 after cleaning the text.

Table 3: Facebook Data Details

Measurements	Counts
Total number of Facebook users	250
Total number of Status before preprocessing	9917
Total number of Status after cleanup (removed status with less than 4 words)	7687
Total words	146159
Total words after preprocessing	64187

Total 29 users tweets scraped for year 2018, Table 4 provided total Tweets count before cleaning the data, and after cleaning the data using Natural language techniques. Tweets removed if a number of words less than 4 after cleaning the text.

Table 4: Twitter Data Details

Measurements	Counts
Total number of Twitter users	29
Total number of Tweets before preprocessing	258301
Total number of Tweets after cleanup (removed tweets with less than 4 words)	218629
Total words	5302073
Total words after preprocessing	1968660

4.3.1 Evaluation Metrics

Evaluated results using different metrics Mean squared error, Precision, Recall, F1-score [19], Average precision score [21].

i. Mean squared error

Mean squared error (MSE), which is the mean squared difference between predicted and actual values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

MSE is the mean of the square of the difference of actual and prediction value

ii. Precision

In binary classification, precision (also called positive predictive value) is the fraction of related instances among the retrieved instances.

$$\frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Precision [19] is used to determine when the cost of a false positive is high. Precision is the ability of classifier not to label as positive a sample which is negative [20].

iii. Recall

In binary classification, recall (also known as sensitivity) [19] is the ratio of correctly identified instances over the total amount of relevant instances.

$$\frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Recall helps to determine when the cost of false negative is high [19]. The recall is the ability of the classifier to find all negative samples [20].

iv. F1-Score

F1-score considered one of the most popular performance metrics. It is also called a balanced F1-score or F-measure. It is a harmonic mean of recall and precision [19]. It is used to test accuracy. It is the consideration of both precision and recall. F1-score considered perfect when the value is 1 and considered as a complete failure when the value is 0. Precision and Recall contributed to the F1-score equally.

$$\frac{2\text{xPrecisionxRecall}}{\text{Precision} + \text{Recall}}$$

v. Average_precision_score

Average_precision_score summarizes [21] a precision-recall curve as the weighted mean of precisions achieved(P) at each threshold, with the increase in recall (R) from the previous threshold used as the weight.

$$AP = \sum_{n} (R_n - R_{n-1}) P_n$$

4.3 Results

Multi-label classification models measurement results compared with different vectorization methods and different machine learning algorithms. Multinomial Naïve Bayes is conditional independence where feature probabilities are independent given the class *c.* Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name logistic regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. A Support Vector Machine (SVM) performs classification by finding the hyperplane that maximizes the margin between the two classes. The vectors that define the hyperplane are the support vectors. For

Linear SVC kernel is Linear function which finds the hyperplane that maximizes the margin and minimizes the misclassifications

4.4.1 Personality Classification

Built personality classification using myPersonality data, myPersonality data collected using Facebook app users participated in psychological research, myPersonality is multi labeled data each status text is labeled with y/n for each personality of Big 5 model. Trained 5 different models with multiple algorithms Multinomial Naïve Bayes[15], Logistic Regression[16] and Linear SVC[17] with TFIDF vectors. Each personality Mean square error compared as follows:

Table 5: Mean Squared Error Comparison of Different Classification Models

Personality Trait	**Multinomial Naive Bayes MSE**	**Logistic SVC MSE**	**Logistic Regression MSE**	**SVM[2]MSE**
cOPN	**0.2483**	0.2719	0.2455	0.5572
cCON	**0.3841**	0.4124	0.4037	0.4477
cEXT	0.3833	0.3986	0.3931	**0.3316**
cAGR	**0.4053**	0.412	0.4214	0.7084
cNEU	**0.3683**	0.3915	0.3785	0.53

Each text is multi labeled for each personality. Built multi-label classification wrapping Multinomial Naïve Bayes[15], Logistic Regression[16], Linear SVC[17] into OneVsRestClassifier[18]. To achieve multi-label classification labels transformed using MultiLabelBinarizer will give results as [1 0 0 1 1]

Table 6: Multi-label Classification for Facebook Data

Multi-label Classification Models	**Precision**	**Recall**	**F1-Score**	**Average Precision Score**
Word2Vec -Multinomial Naïve Bayes	0.64	0.51	0.56	0.58
Word2Vec -Logistic Regression	0.64	0.51	0.56	0.58
Word2Vec -Linear SVC	0.64	0.51	0.56	0.58
TFIDFVectorizer -Multinomial Naïve Bayes	**0.67**	**0.6**	**0.63**	**0.6**
TFIDFVectorizer -Logistic Regression	0.65	0.6	0.63	0.6
TFIDFVectorizer -Linear SVC	0.63	0.63	0.63	0.59
LDA Features -Multinomial Naive Bayes	0.64	0.5	0.56	0.57
LDA Features -Logistic Regression	0.64	0.49	0.56	0.58
LDA Features -Linear SVC	0.64	0.49	0.56	0.58

For Facebook data Multinomial Naïve Bayes [15] with TFIDFVectorizer got high measurement scores precision, recall, F-1 score compared to other models given in Table 6. models. The average Precision score with TFIDF vector Multinomial Naïve Bayes [15] is 60 %. Average precision is a measure that combines recall and precision for ranked retrieval results. For one information need, the average precision is the mean of the precision scores after each relevant document is retrieved.

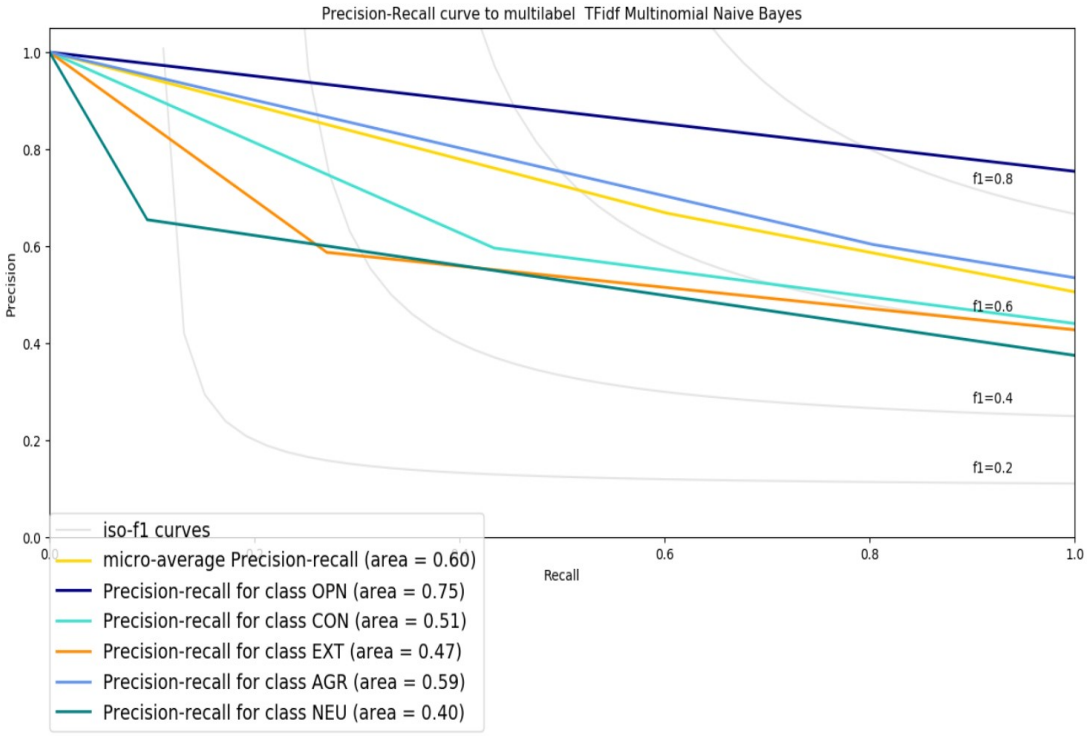

Figure 5: Facebook Multi-label Classification Average Precision

Figure 5 provides the overall average precession and Precision recall for each personality. openness Precision-recall is 75% and neuroticism precision-recall value is 40%. This is due to the number of Facebook status high for the openness personality vs. neuroticism. Using the Facebook model labeled Twitter data, Table 7 provides the precision, recall and F1-score measurements for the Twitter model with different algorithms.

Table 7: Multi-label Classification for Twitter Data

Multi-label Classification Models	**Precision**	**Recall**	**F1-Score**	**Average Precision Score**
Word2Vec -Logistic Regression	0.83	0.73	0.78	0.74
Word2Vec -Linear SVC	0.83	0.73	0.78	0.74
TFIDFVectorizer -Multinomial Naïve Bayes	0.88	0.79	0.83	0.8
TFIDFVectorizer -Logistic Regression	0.95	0.91	0.93	0.92
TFIDFVectorizer -Linear SVC	**0.96**	**0.95**	**0.96**	**0.95**
LDA Features -Multinomial Naïve Bayes	0.93	0.91	0.9	0.87
LDA Features -Logistic Regression	0.92	0.91	0.9	0.86
LDA Features -Linear SVC	0.93	0.92	0.92	0.87

For Twitter data, LinearSVC with TFIDFVectorizer got high measurement scores compared to other models. Average precision for Neuroticism is low as the data size for the neuroticism is lower compared to other personality.

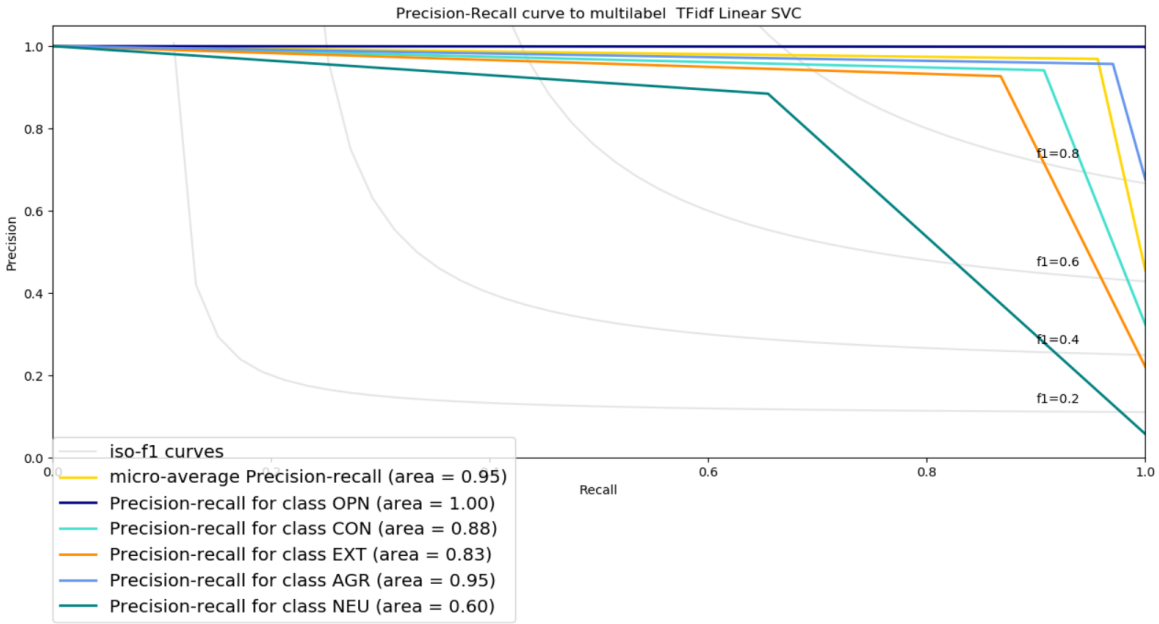

Figure 6: Twitter Multi-label Classification Average Precision

Twitter is the social media where people are more open. Table 8 shows personality of the tweets for every user tweet for the year of 2018. All the users first high personality is Openness and the second highest is Agreeableness. Figure 7 shows President Trump tweets counts for each personality

Table 8: Tweets Classification Details for Each User

User	OPN	CON	EXT	AGR	NEU	No of Tweets
0	5865	1651	1228	4028	339	64625
1	9758	3685	2647	6002	459	107635
2	14233	4983	3049	9261	1003	156882
3	2901	1002	575	2287	94	31955
4	10719	2842	2371	7223	703	118250
5	29521	10921	7984	20522	2131	325974
6	2086	637	390	1434	82	22957
7	13965	4259	3146	8855	829	154000
8	714	222	134	499	35	7854
9	484	168	82	354	28	5324
10	952	241	177	659	48	10483
11	7566	2037	1620	4554	444	83512
12	35160	11105	7088	25353	1716	387211
13	1402	539	440	925	59	15444
14	3878	1036	655	2360	234	42724
15	2597	836	451	1736	133	28600
16	42806	13964	9568	29174	2702	472912
17	4426	1476	1061	2719	242	48785
18	5123	1983	1187	3522	231	56463
19	5509	1435	962	3665	428	60665
20	1905	579	375	1328	107	20955
21	379	117	89	286	15	4180
22	1074	337	243	755	66	11836
23	1893	528	337	1353	95	20856
24	1111	220	181	836	80	12265
25	4159	1285	660	3304	158	45782
26	4554	1280	1074	3267	159	50149
27	3325	1229	903	2109	163	36641
28	2797	1304	741	1883	140	30767

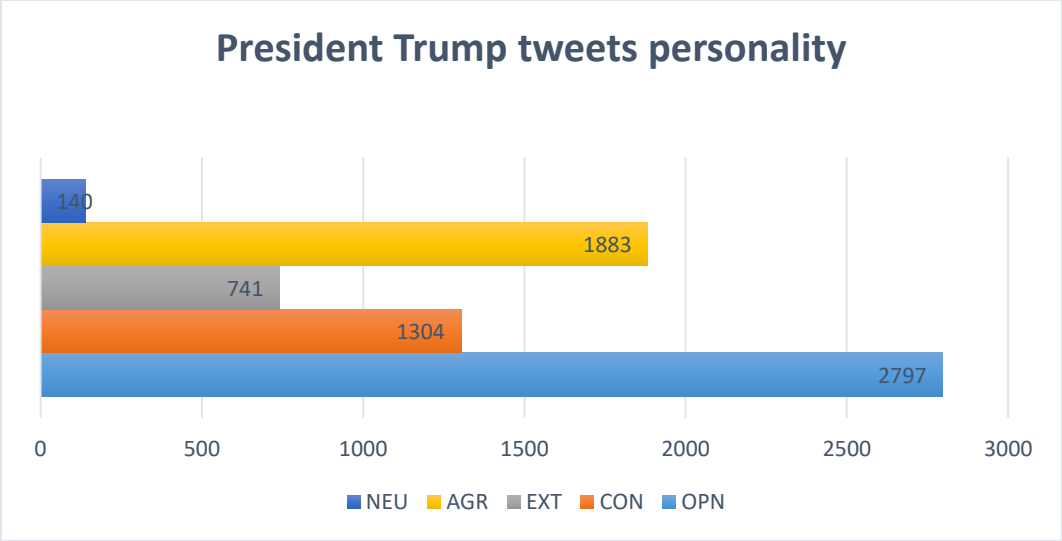

Figure 7: President Trump Tweets Personality

4.4.2 Personality Topic Terms Results

Both Facebook and Twitter are small and not connected documents. Identified the Topic terms for each personality based on the definition of the personality description.

People with Openness are Imaginative, Preference for variety, Independent. Table 9 provides some of the interesting topic terms for the Facebook, Twitter model provides openness personality talk about people, arts, exercise, vacation, and industry. Topic terms show openness people are very expressive

Table 9: Important Topic Terms for Openness

Topic Model	Openness Topic Terms
Facebook Model	children, sister art, poem, sing, song vacation, vegas, squats, walking, skateboarding beautiful, loves, hoping, enjoying thesis
Twitter Model	mother, baby, students, kids, woman, man, people congratulations, wow, good, happy, pretty, great live, love, life amazon, google, Netflix

People with Conscientiousness are Organized, Careful, Disciplined. Table 10 has some of the interesting topic terms for Facebook, Twitter model provides Conscientiousness personality discussed about health, food, time and places. From the topic words, it shows people who are conscientious looks to be worried.

Table 10: Important Topic Terms for Conscientiousness

Topic Model	Conscientiousness Topic Terms
Facebook Model	thoughts, feelin, pretty Sick, die , love Heartbreaking, disorganized, terrified, failure, incongruous Burger, cheesecake Jewish
Twitter Model	Florida, India, Georgia, London, Hawaii. ohio American, Canadian eagles, sea, wind, earth, ,air Christmas July, hour, minutes falling , dying

People with Extraversion are Sociable, Fun Loving, Affectionate. Table 11 describes some of the interesting topic terms for Facebook/Twitter model provides extraversion personality, extraversion people looks to be very appreciative, like events.

Table 11: Important Topic terms for Extraversion

Topic Model	Extraversion Topic Terms
Facebook Model	Appreciates kind, coolest, cheer, amazingly wedding litomysl(Historical place in Germony) social endorphin (workout)
Twitter Model	lol, amazing, inspiring, laugh, futuristic, pleasure concerts, hollywood, jongleurs

People with Agreeableness are Self-hearted, trusting, helpful. Table 12 describes some of the interesting topic terms for Facebook/Twitter model provides Agreeableness personality, topic words show they care about friends, family, happy when they are expressing in social media and enjoy partying.

Table 12: Important Topic terms for Agreeableness

Topic Model	**Agreeableness Topic Terms**
Facebook Model	Believes wow, fun temples life coworkers, family, mom partying, adventurous, canoeing cooking
Twitter Model	Kind listen healthy, beautiful, pretty, super artist women, student, girl care, home

People with Neuroticism are Anxious, Insecure, Self-pitying. Table 13 got some of the interesting topic terms for Facebook/Twitter model provides Neuroticism personality, topic words showing people with neuroticism personality worried about health, crime and more talking about risk

Table 13: Important Topic terms for Neuroticism

Topic Model	Neuroticism Topic Terms
Facebook Model	horoscope, summer rumors sudden, obsessed, regret, confusing, losing grave divine cannibalism
Twitter Model	concerned, fear, upset, mess, annoying ich, hydrocortisone, disease risks, disaster criminals, prison republicans

President Trump tweets for Neuroticism are very low to find out the topic terms. President tweets contains terms related to companies for the Conscientiousness.

Table 14: President Trump Tweets Important Topic Words

Openness	Conscientiousness	Extraversion	Agreeableness
gulfport	companies	thrilled	Women
championship	manufacturing	mexico	Pakistan
encourage	google	honduras	immigration
loved	arrest	legendary	forced
incredibly	terrorist	honored	captured
wall	losing	globalist	guilty
concrete	prayers	civil	progress
farmers	worry	army	

4.4.3 Personality Based Topic Interests

Figures 8-12 provides interested topics for each personality. Figure 8 shows personality with openness, extraversion, neuroticism interested in sports football, tennis. Figure 9 shows all the personalities interested in politics. Figure 10 shows personality openness, conscientiousness, extraversion, neuroticism interested in technology. Figure 11 and Figure 12 shows all of the personalities interested in Money and Religion.

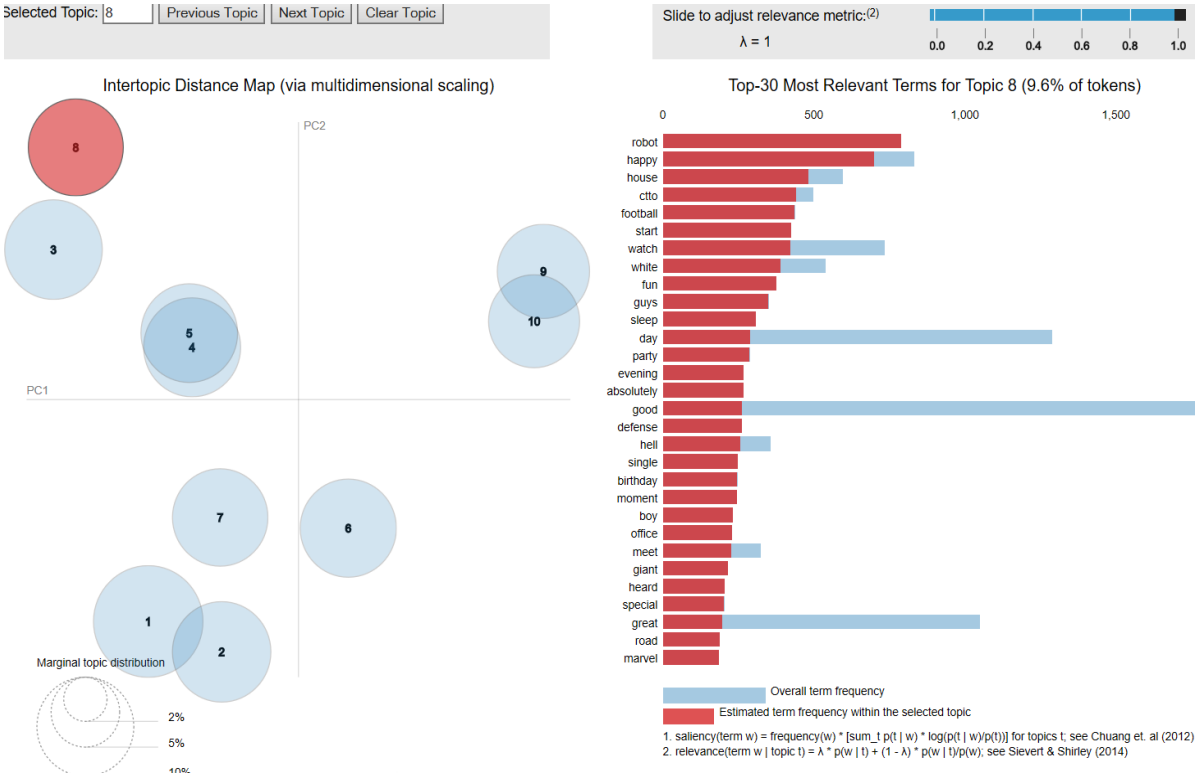

Figure 8: Sports - Openness, Extraversion, Neuroticism Interested

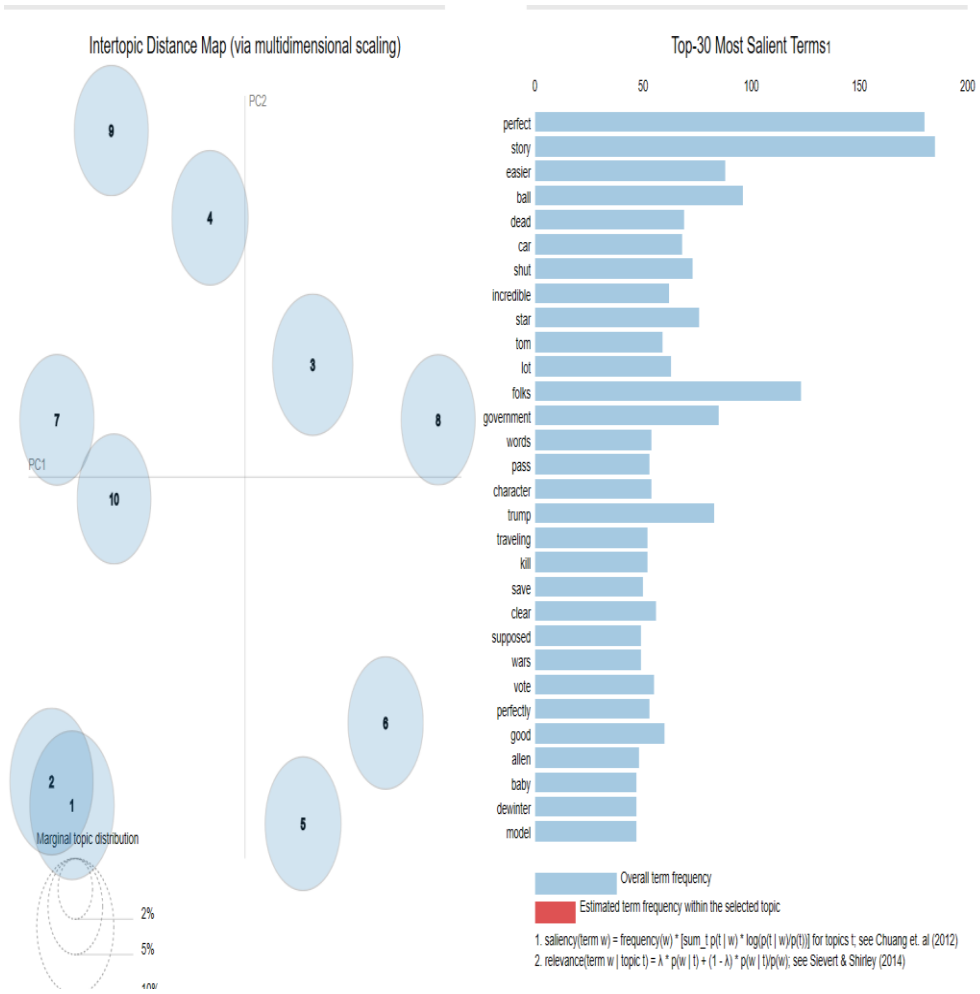

Figure 9: Politics - All of the Personalities Interested

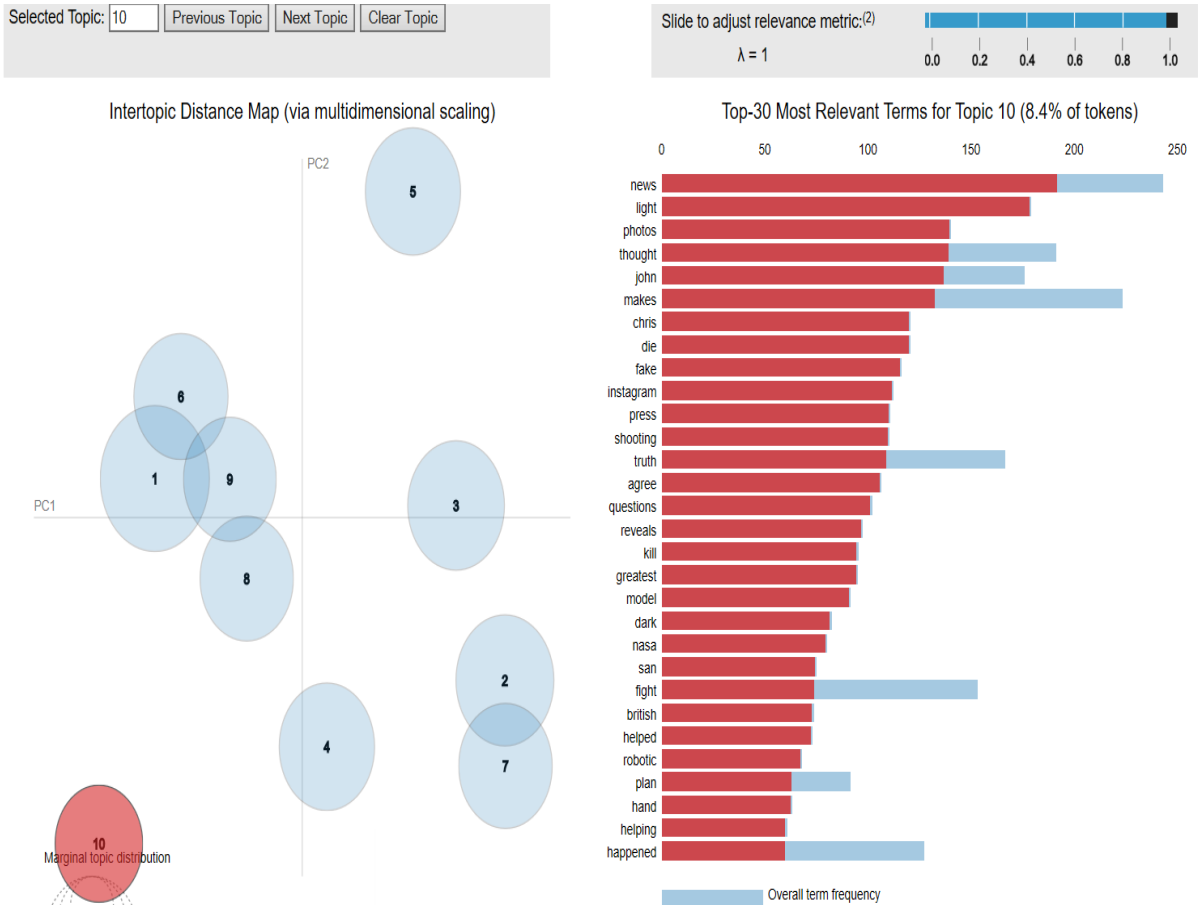

Figure 10: Technology - Openness, Conscientiousness, Extraversion, Neuroticism

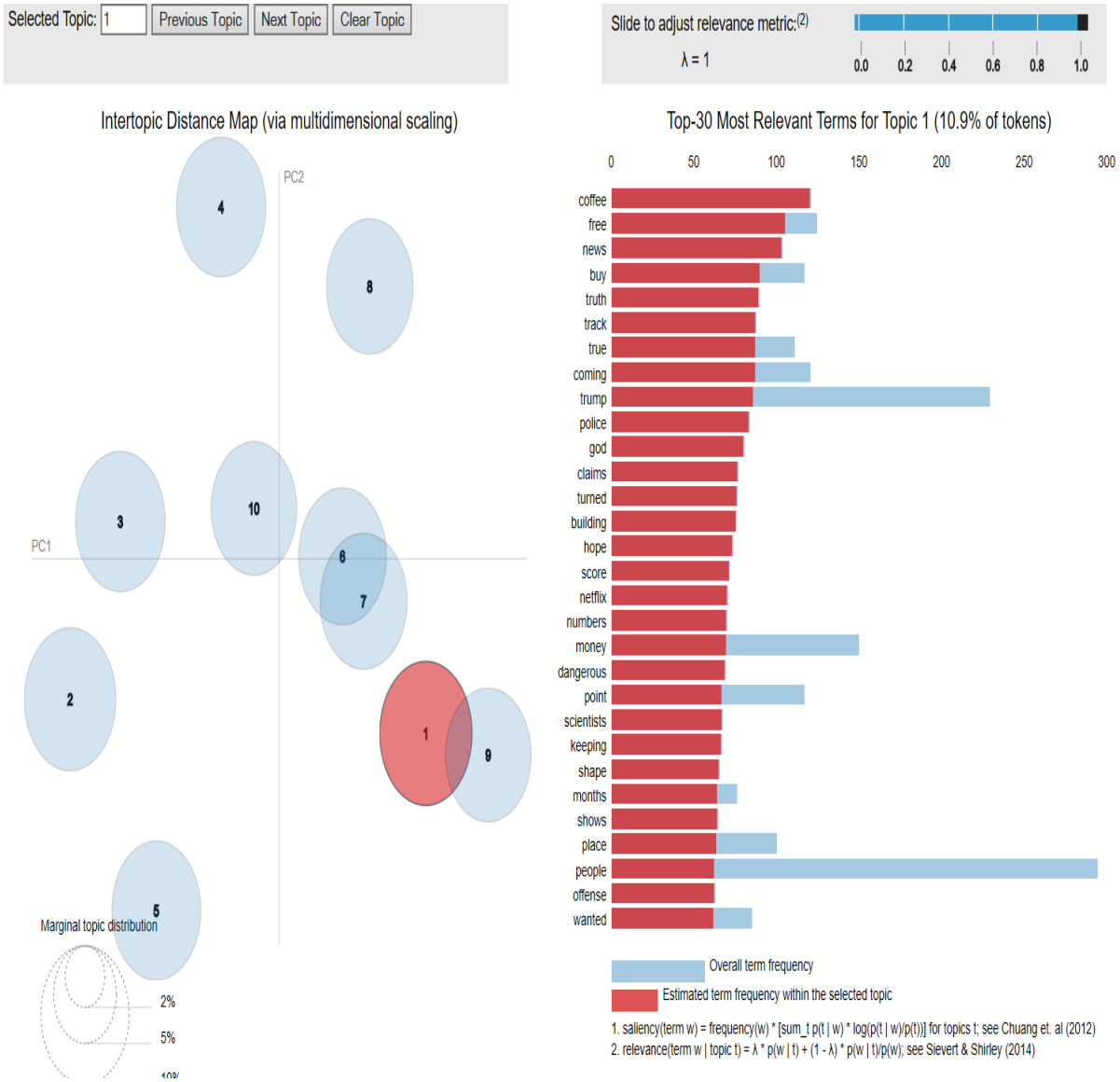

Figure 11: Money – All of the Personalities Interested

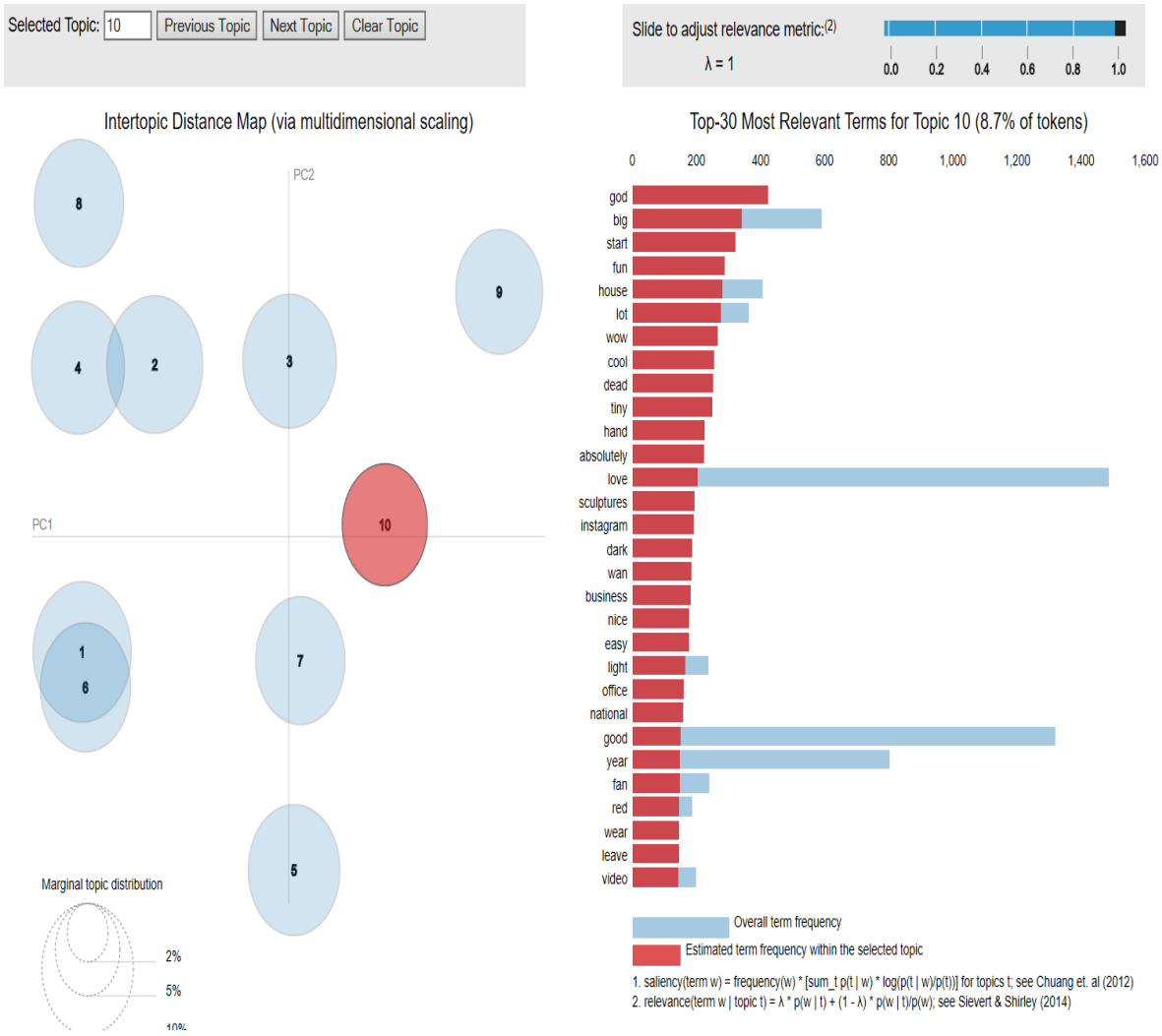

Figure 12: Religion – All of the Personalities Interested

CHAPTER 5. CONCLUSION AND FUTURE WORK

5.1 Conclusion

In this thesis, we presented the implementation of the multi-label classification. Facebook myPersonality data which is labeled using the Big 5 personality model, Facebook data used to train models using Multinomial Naïve Bayes [15], Logistic Regression [16], Linear SVC [17] machine learning models. Twitter data for random 28 users and president trump data scraped for the year 2018, labeled tweets for multiple personalities using the Facebook model. Wrapping Multinomial Naïve Bayes[15], Logistic Regression[16], Linear SVC[17] into One vs Rest classification with TFIDFVectorizer, Word2Vec vector size 100 using the Pipeline approach achieved multi-label classification. Multi-nominal Naïve Bayes performed better for Facebook data compare to other models. Linear SVC provided 95 precision for Twitter data. LDA [1] provided good topic terms closed to personality qualities. KATE [2] provided topics associated with each personality. Openness, Extraversion, Neuroticism interested in Sports. Openness, Conscientiousness, Extraversion, Neuroticism interested in Technology. All the personalities were interested in Politics, Money and Religion.

5.2 Future Work

Data for each personality counts is not balanced. Apply the models with balanced data for each personality the same count build multi-label models and compare the results. Sentiment analysis can be done on the Text data. Extend the personality classification model with sentiment analysis to find out if there is any relation between the personality traits and sentiment analysis. Personality classification can be extended to other social media data like Instagram. Depression suicides are the biggest problem in the society extend the research to find the relation between the personality and depression identify the text related to depression in social media

BIBLIOGRAPHY

[1] D. M. Blei , A. Y. Ng and M. Jordan, "Latent Dirichlet Allocation", *Journal of Machine Learning Research* , Vol 3, no. 4-5, 993-1022, 2003, DOI : 10.1162/jmlr.2003.3.4-5.993

[2] Y. Chen, and M. J. Zaki. "KATE: k-competitive autoencoder for Text." 2017, https://arxiv.org/pdf/1705.02033.pdf

[3] S. V. Paunonen and M. C. Ashton "Big five factors and facets and the prediction of behavior." *Journal of Personality and Social Psychology, vol. 81, no.* 3, 524–539 , 2001

[4] L. W. McCray, H. R. Bogner, M. D. Sammel, ScD and J. J. Gallo, "The role of patient personality in the identification of depression in older primary care patients*." Int J Geriatr Psychiatry*. 2007 Nov; Vol. 22, no. 11: 1095–1100, DOI :10.1002/gps.1791

[5] Wikipedia, "Twitter" , Internet: https://en.wikipedia.org/wiki/Twitter

[6] F. Celli, F. Pianesi, F, D. Stillwell and M. Kosinski "Workshop on computational personality recognition: shared task." , 2013*, Seventh International AAAI Conference on Weblogs and Social Media. Computational Personality Recognition*.

[7] J. Golbeck, C. Robles, M. Edmondson, and K. Turner. "Predicting personality from Twitter.", 2011, in *IEEE Third International Conference on and 2011 IEEE Third International Conference on Social Computing*

[8] G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio. "TwitPersonality: computing personality traits from Tweets using word embeddings and supervised learning." *Information* (Switzerland), vol. 9, no. 5, 2018. DOI: 10.3390/info9050127

[9] N. Majumder, S. Poria, A. Gelbukh and E. Cambria. "Deep learning based document modeling for personality etection from Text." *IEEE -Intelligent Systems*,Vol. 32, no. 2 , 2017, DOI: 10.1109/MIS.2017.23

[10] W. Li, Y. Chen, T. Hu, J. Luo. "Mining the relationship between emoji usage patterns and personality." , 2018, *Twelfth International AAAI Conference on Web and Social Media* https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/viewPaper/17830

[11] N. Ahmad and J. Siddique "Personality assessment using Twitter tweets." Vol. 112, 2017, DOI: 10.1016/j.procs.2017.08.067

[12] D. Alvarez-Melis and M. Saveski "Topic modeling in Twitter: aggregating tweets by conversations." *Proceedings of the Tenth Inter-national AAAI Conference on Web and Social Media*, 2016 https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/download/13162/12778

[13] A. F. Hidayatullah, E. C. Pembrani, W. Kurniawan and G. Akbar, R. Pranata "Twitter topic modeling on football news." *IEEE 3rd International Conference on Computer and Communication Systems 2018*

[14] B. Baharudin , L. H. Lee, K. Khan and A. Khan, "A Review of Machine Learning Algorithms for Text-Documents Classification", *Journal of Advances in Information Technology,* Vol. 1, 2010, DOI : 10.4304/jait.1.1.4-20

[15] S. Xu, Y. Li and W. Zheng, "Bayesian Multinomial Naïve Bayes Classifier to Text Classification" , 2017, DOI: 10.1007/978-981-10-5041-1_57

[16] NCSS Statistical Software, "Logistic Regression" , Internet: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Logistic_Regression.pdf

[17] S. Sayad, "Linear SVC", Internet: https://www.saedsayad.com/support_vector_machine.htm

[18] scikit-learn.org, "sklearn.multiclass.OneVsRestClassifier scikit-learn 0.22.1" ,Internet: https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html

[19] Pathmind "A.I Wiki" , Internet: https://skymind.ai/wiki/accuracy-precision-recall-f1

[20] scikit-learn.org,"sklearn.metrics.precision_recall_fscore_support — scikit-learn 0.22" ,Internet: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html

[21] scikit-learn.org, "sklearn.metrics.average_precision_score — scikit-learn 0.22" ,Internet: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html

[22] Dreamtalent ,"The Best Psychometric Testing tools for recruitment". Internet : https://dreamtalent.id/blog/stop-using-mbti-disc-theyre-not-that-good-20190508025206

[23] D. Schmitt, J. Allik, R. McCrae, and V. Benet-Martinez. "The geographic distribution of Big five personality traits: patterns and profiles of human self-description across 56 nations." in *Journal of Cross-Cultural Psychology*, vol. 38, no. 2, 173-212, 2007.

[24] A. F. Hidayatullah and M. F. Ma'arif. "Road traffic topic modeling on Twitter using Latent Dirichlet Allocation." *IEEE International Conference on Sustainable Information Engineering and Technology (SIET)* 2017, DOI: 10.1109/SIET.2017.8304107

[25] M. Hagras,G. Hassan and N. Farag. "Towards natural disasters detection from Twitter using topic modelling." *European Conference on Electrical Engineering and Computer Science (EECS)*, 2017, DOI: 10.1109/EECS.2017.57

[26] J.A. Johnson, "Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120." *Journal of Research in Personality*, Vol. 51, DOI: 10.1016/j.jrp.2014.05.003

[27] "Multi lable data - Questions and Answers from Cross Validated, the statistics and machine learning QA site from the Stack Exchange". Internet: https://www.kaggle.com/stackoverflow/statsquestions

[28] Y. Liu, J. Wang , Y. Jiang. "PT-LDA:A latent variable model to predict personality traits of social network users" *article in Neurocomputing* , Vol. 210,2016, DOI: 10.1016/j.neucom.2015.10.144

[29] J. W. Pennebaker, and L. A. King "Linguistic styles: Language use as an individual difference." *in Journal of Personality and Social Psychology*, vol. 77, no. 6, 1296–1312, 2000, DOI: 10.1037//0022-3514.77.6.1296

[30] N. V. Ven, A. Bogaert, A. Serlie, M. J. Brandt and J. J.A. Denissen "Personality perception based on LinkedIn profiles" , *Journal of Managerial Psychology,* Vol. 32, no. 6, 418-429, 2017 DOI: 10.1108/JMP-07-2016-0220

[31] A. Jusupovam, F. Batista and R. Ribeiro "Characterizing the Personality of Twitter Users based on their Timeline Information.", DOI : 10.18803/capsi.v16.292-299

[32] M. Selfhout, W. Burk, S. Branje, J. Denissen, M. van Aken, and W. Meeus. "Emerging Late Adolescent Friendship Networks and Big Five Personality Traits: A Social Network Approach.", *Journal of personality*, Vol. 78, no. 2, 509–538, 2010. DOI: 10.1111/j.1467-6494.2010.00625.x

[33] A. Ng, M. W. Bos, L. Sigal, B. Li , "Predicting Personality from Book Preferences with User-Generated Content Labels ", *Article in IEEE Transactions on Affective Computing*, 2017, DOI: 10.1109/TAFFC.2018.2808349

[34] M. Carbonneau, E. Granger, Y. Attabi and G. Gagnon , "Feature Learning from Spectrograms for Assessment of Personality Traits", *Article in* in *IEEE Transactions on Affective Computing,* 2016, DOI : 10.1109/TAFFC.2017.2763132

VITA

Trinadha R Muppala completed her Master's in Computer Applications from Motilal Nehru National Institute of Technology Allahabad, India. She started her Masters in Computer Science at the University of Missouri-Kansas City (UMKC) in August 2017, with an emphasis on Data Sciences and graduates in December 2019. Trinadha R Muppala working as Lead Architect Logistics at Sprint IT. She leads Order Inventory Management and Retail applications at Sprint. She leads group of developers at Sprint. She also volunteers for Asha for Education from 2009. Asha for Education is a fully volunteer-run 501(c)(3) non-profit organization with 50+ chapters around the world whose mission is to catalyze socio-economic change in India through the education of underprivileged children. She is a project steward for Community/home-based rehabilitation of Children with Intellectual & Developmental Disabilities.