Inferring Big 5 Personality from Online Social Networks

Geetha Sitaraman

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

University of Washington

2014

Committee:

Martine De Cock, Chair

Sergio Davalos

Ankur Teredesai

Golnoosh Farnadi

Shanu Sushmita

Program Authorized to Offer Degree:
Computer Science and Systems

University of Washington

Abstract

Inferring Big 5 Personality from Online Social Networks

Geetha Sitaraman

Chair of the Supervisory Committee:
Associate Professor Martine De Cock
Institute of Technology
University of Washington

Online social networks are very popular with millions of people creating online profiles and sharing personal information including their interests, activities, likes/dislikes and thoughts with their friends and family. This rich user generated content from social media makes them an ideal platform to study human behaviour. In our research, we are interested in latent variables such as the long term personality traits and the short term emotional state of users. Proper mining of the user generated content can be used to identify personality traits of users without having them fill out questionnaires. These traits are shown to strongly influence a person's decisions, behavior and preferences for language, music, books etc. We explore the use of different machine learning techniques and feature selection methodologies for inferring users' personality traits using information available from their online profile. We study five multivariate regression algorithms and contrast them with a single target approach for predicting the scores. Additionally, we explore feature subset selection using correlation based heuristics and evaluate the quality of the feature space produced using two different machine learning algorithms: Linear Regression and Support Vector Regressors. The performance of the above techniques is evaluated on two different datasets: a myPersonality dataset collected from Facebook and a YouTube personality dataset collected from video posts of vloggers. All five multivariate as well as single target algorithms and correlation based feature selection methods outperformed the average baseline model for all

five personality traits on both the datasets. Furthermore, we study the relation between emotions expressed in approximately 1 million Facebook (FB) status updates and the users' personality, age, gender and time of posting. We use this in establishing associations such as open personality users express emotions more frequently, while neurotic users are more reserved. With the ability to identify users' personality and emotions, advertisements could be tailored based on the user's personality type since personality and/or emotion-aware interfaces are more persuasive.

TABLE OF CONTENTS

Page

List of Figures . . . iii

List of Tables . . . iv

Chapter 1: Introduction . . . 1
- 1.1 Online Social Networks . . . 1
- 1.2 Personality-Based User Modelling . . . 1
- 1.3 Big 5 Personality . . . 2
- 1.4 Problem Statement . . . 4
- 1.5 Contribution . . . 4

Chapter 2: Background . . . 6
- 2.1 Regression Model for Personality Recognition . . . 7
- 2.2 Evaluation . . . 12
- 2.3 Feature Selection Methodology . . . 14

Chapter 3: Datasets Description and Preparation . . . 18
- 3.1 myPersonality Dataset . . . 18
- 3.2 myPersonality Dataset - Features Used . . . 20
- 3.3 YouTube vloggers dataset . . . 23
- 3.4 YouTube vloggers Dataset - Features Used . . . 24

Chapter 4: Emotions and Personality in Facebook . . . 29
- 4.1 Introduction and Motivation . . . 29
- 4.2 Emotion Detection . . . 30
- 4.3 Results Discussion . . . 31
- 4.4 Conclusions and Future Work . . . 36

Chapter 5: Regression Approach to Big 5 Personality Recognition . . . 38
- 5.1 Experimental Setup . . . 38

5.2 Experiments on YouTube vloggers dataset 39
5.3 Experiments on myPersonality Dataset 47
5.4 Conclusions and Future Work 52

Chapter 6: Conclusion 55
6.1 Summary 55
6.2 Future Work 56

Bibliography 58

Appendix A: Correlations for Youtube VLoggers Dataset 65

Appendix B: Correlations for MyPersonality Dataset 71

Appendix C: Correlations between emotions and Facebook features 76

LIST OF FIGURES

Figure Number Page

2.1 Soft Margin loss setting in linear SVM [65] 9

3.1 Distribution of the Big 5 Scores - Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticm scores among the 38,106 users. 21

3.2 An example of an excerpt from a vlogger transcript 24

3.3 Distribution of the perceived Big 5 Scores - Openness, Conscientiousness, Extraversion, Agreeableness, Emotional Stability scores among the 404 vloggers. 25

4.1 Emotion frequency in Facebook status updates. Almost 60% of the status updates express at least one kind of emotion from the NRC lexicon, and many posts convey more than one emotion. 31

4.2 Probability of occurrences of emotions in status updates from female and male users. 32

4.3 Probability of occurrences of emotions in status updates from users of different age groups. Users are more likely to post emotions as they get older. . . . 33

4.4 Probability of occurrences of emotions in status updates depending on the day of the week. Status updates are more likely to contain emotions during workdays than during the weekend. 34

4.5 Probability of occurrences of emotions in status updates depending on the day of the week. Status updates are more likely to contain emotions during workdays than during the weekend. 35

4.6 Probability of occurrences of emotions in status updates from users of different age groups. Users are more likely to post emotions as they get older. . . . 35

4.7 Probability of occurrences of emotions in status updates from users of different age groups. Users are more likely to post emotions as they get older. . . . 36

LIST OF TABLES

Table Number Page

2.1 Pearson product-moment correlation results among personality scores on 5 traits: *Extraversion (Extr)*, *Agreeableness (Agr)*, *Conscientiousness (Cons)*, *Emotional Stability (EmoStab)*, *Openness (Open)* on YouTube vloggers dataset. Significant correlations ($p < .05$) among the personality scores are indicated in bold. 11

2.2 Pearson product-moment correlation results among personality scores on 5 traits: *Extraversion (Extr)*, *Agreeableness (Agr)*, *Conscientiousness (Cons)*, *Neuroticm (Neu)*, *Openness (Open)* on myPersonality dataset. Significant correlations ($p < .05$) among the personality scores are indicated in bold. 11

3.1 myPersonality data statistics 19

3.2 (Table on the left) Characteristics of female and male users in the dataset. The entire dataset contains 969,035 status updates written by 5,865 users. (Table on the right) Score threshold and number of users for each personality trait. Note that the same user can exhibit more than one personality trait at once. 20

3.3 (Table on the left) Characteristics of 38,106 users in the dataset. (Table on the right) Mean and Standard Deviation of Big 5 personality scores of 38,106 users. 22

3.4 (Table on the left) Characteristics of 404 users in the YouTube vloggers dataset. (Table on the right) Mean and Standard Deviation of perceived Big 5 personality scores of the users. 26

5.1 Forward search: Feature set category and the corresponding letter codes. 41

5.2 RMSE Comparison on #348 training on a YouTube vloggers dataset using all and correlated features under each feature set category. All results are based on 10 fold cross validation using SVM (rbf kernel, C=1). 43

5.3 Best feature set combination obtained using Forward Search on the #348 training on a YouTube vloggers dataset for the 5 traits: *Extraversion (Extr)*, *Agreeableness (Agr)*, *Conscientiousness (Cons)*, *Emotional Stability (EmoStab)*, *Openness (Open)*. All results are based on 10 fold cross validation using SVM (rbf kernel, C=1) 43

5.4 Root mean square error ($RMSE$) and Coefficient of determination (R^2) results for personality trait prediction using *univariate* and *multivariate* regression algorithms on a YouTube vloggers dataset. In each column, the lowest error and highest determination are typeset in bold. Significant differences ($p < .05$) are marked using $*$. 45

5.5 Root mean square error (RMSE) results for personality trait prediction on #38,106 users of myPersonality dataset using all features under each feature set. All values are averaged over 10 fold cross validation using linear regression. In each column, the lowest error are typeset in bold. 50

5.6 Root mean square error ($RMSE$) and Coefficient of determination (R^2) results for the personality trait prediction using *univariate* linear regression and *multivariate* regression algorithms on #38,106 and #3400 users respectively on a myPersonality dataset. In each column, the lowest error and highest determination are typeset in bold. 51

5.7 Comparison of the best model for the 5 traits: *Extraversion (Extr)*, *Agreeableness (Agr)*, *Conscientiousness (Cons)*, *Emotional Stability (EmoStab)*, *Openness (Open)* on the myPersonality and the YouTube vloggers dataset. Each row consists of algorithm, feature set used and the obtained RMSE improvements measure from the mean baseline. See Section 5.2 and Section 5.3 for details. 53

A.1 (Right) Correlation results between the emotional features and the 5 personality impression scores. (Left) Correlation results between the MRC features and the 5 personality impression scores. . . 65

A.2 Correlation results between gender and the 5 personality impression scores. The absolute value of Spearman's rank is reported here. 65

A.3 Correlation results between the LIWC features and the 5 personality impression scores. (Right) 1-25 Features. (Left) 26-51 Features. 66

A.4 Correlation results between the LIWC features and the 5 personality impression scores. 52-81 Features. 67

A.5 Correlation results between the audio-video features and the 5 personality impression scores. . . . 68

A.6 Correlation results between the Splice features and the 5 personality impression scores. 1-37 Features. 69

A.7 Correlation results between the Splice features and the 5 personality impression scores. 38-74 Features. 70

B.1 Correlation results between the Demographic and Facebook Activity features and the 5 personality scores. 71

B.2 Correlation results between the LIWC features and the 5 personality impression scores. (Right) 1-25 Features. (Left) 26-51 Features. 72

B.3 Correlation results between the LIWC features and the 5 personality impression scores. 52-81 Features. 73

B.4 Correlation results between the Splice features and the 5 personality impression scores. 1-32 Features. 74

B.5 Correlation results between the Splice features and the 5 personality impression scores. 33-71 Features. 75

C.1 Pearson Chi Squared test results for on characteristics of users and posts, and emotion categories: *positive (Pos)*, *negative (Neg)*, *anger (Ang)*, *anticipation (Ant)*, *disgust (Dis)*, *fear (Fea)*, *joy (Joy)*, *sadness (Sad)*, *surprise (Sur)*, and *trust (Tru)*. 77

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Prof. Martine De Cock for the continued help and advice she has given me throughout our work. Her constructive remarks and comments helped a great deal in advancing our work. I received a lot of guidance from my mentor, Golnoosh Farnadi who is always willing to share her ideas in every aspect of our work. Her enthusiasm and hard work has always inspired me. I would like to thank her for all that she has done for the success of our work. I am thankful to Prof. Sergio Davalos, who has always brought a different perspective to our work through his varied experience and knowledge. His remarks enhanced the quality of our work. Special thanks to Dr. Shanu Sushmita for the thoughful suggestions on many aspects that had helped me develop the work to what it is now. My thanks to Prof. Ankur Teredesai for setting me in this way towards my goal by motivating me to take up this work. I would also like to thank my fellow student, Nhat for sharing her knowledge on several softwares that helped me in learning them quickly. Special thanks to David Hazel for providing us with the needed resources for being able to conduct our research.

Finally, I am thankful to my family including my mother, Subhalakshmi, my husband, Shiva and my kids, Shripadh and Shriya for their endless patience, support and confidence in me. I am truly grateful for my family for making all the sacrifices while I was in school.

DEDICATION

To my family

Chapter 1

INTRODUCTION

1.1 Online Social Networks

Online social networks have become very popular nowadays with an ever increasing number of people using them everyday. A social network can be viewed as a structure comprising of actors and the connections that exist among them. In the context of online social media, a social network is made up of individuals and connections (friendship, family relation, following) between them. Among many other reasons, people use online platforms to stay connected with their friends and famliy and to be able to share and express ideas. Companies such as Facebook, LinkedIn, Twitter and Google specialise in this core concept but offer their service to the users in slightly different ways. Facebook, for example, is among the most visited websites with a user base of more than 829 million people using it on a daily basis [20]. Users on this social networking site can create an online profile with personal information and have the ablility to invite their friends and colleagues to view their profile and post comments [38]. The huge user base, with its rich user generated content, provides a great opportunity for conducting research on social network analysis and user behavioural modelling. Social media serves as an ideal platform for studying human behaviour since it has been shown that profiles created on social media are a true reflection of the user's personality impression, rather than an idealised self view [6]. User generated content in the form of comments and posts provides insight into user opinions and behaviour which could be used for marketing analytics and product sentiment analysis.

1.2 Personality-Based User Modelling

Interactive and personalised interfaces are ubiquitous in all walks of modern life today. These are systems that adapt to an individual user or group of users' goals, tasks and interests by using the information available about a user [13]. User Modelling is the process of building an internal representation of a user based on the data gathered about the user. Users can be characterised based on their age, gender, interests, preferences, likes and dislikes, etc. Examples of personalised systems include web personalisation [48], recommender systems for ecommerce, news and entertainment systems etc. Various levels of personalisation could exist within a system using explicit user characteristics like age, gender demographics or implicit user behaviour pattern (web browsing history, click pattern).

Social media websites provide a unique opportunity for personalized services to use other aspects of user behavior. Besides users' structured information contained in their profiles, e.g., demographics, users produce large amounts of data about themselves in variety of ways including textual (e.g., status updates, blog posts, comments) or audiovisual content (e.g., uploaded photos and videos). Many latent variables such as personalities, emotions and moods — which, typically, are not explicitly given by users — can be extracted from user generated content (see e.g. [22, 28]). Research into automatic personality prediction is a very nascent area which is gaining increased research attention due to the potential in many computational applications. It is shown that people are more receptive and engaged when computer interfaces and messages are presented from an user's perspective and exhibit similar traits as them [53]. Personality can affect the decision making process and has been shown to be relevant in the selection of music, movies, TV programs and books. It has been shown that personality affects preference for websites [42], language used in online social media [66], choice of Facebook Likes [43], music taste [64], and content such as movies, TV shows, and books [14].

Having the ability to predict personality from social media is very valuable for many applications like employers who wish to evaluate a potential candidate, friend recommendation systems, dating websites for better matching, marketing and advertisement for tailored targeting etc.

1.3 Big 5 Personality

Personality is a fundamental differentiating factor of human behavior. Research in the psychology literature has led to a well established model for personality recognition and description, called the Big Five Personality Model. The five traits can be summarized in the following way [19]:

- **Openness to experience** (Openness) is related to imagination, creativity, curiosity, tolerance, political liberalism, and appreciation for culture. People scoring high on Openness like change, appreciate new and unusual ideas, and have a good sense of aesthetics.

- **Conscientiousness** measures preference for an organized approach to life in contrast to a spontaneous one. People scoring high on Conscientiousness are more likely to be well organized, reliable, and consistent. They enjoy planning, seek achievements, and pursue long-term goals. Non-conscientious individuals are generally more easy-going, spontaneous, and creative. They tend to be more tolerant and less bound by rules and plans.

- **Extroversion** measures a tendency to seek stimulation in the external world, the company of others, and to express positive emotions. People scoring high on Extroversion tend to be more outgoing, friendly, and socially active. They are usually energetic and talkative; they do not mind being at the center of attention, and make new friends more easily. Introverts are more likely to be solitary or reserved and seek environments characterized by lower levels of external stimulation.

- **Agreeableness** relates to a focus on maintaining positive social relations, being friendly, compassionate, and cooperative. People scoring high on Agreeableness people tend to trust others and adapt to their needs. Disagreeable people are more focused on themselves, less likely to compromise, and may be less gullible. They also tend to be less bound by social expectations and conventions, and more assertive.

- **Emotional Stability** (reversely referred to as Neuroticism) measures the tendency to experience mood swings and emotions such as guilt, anger, anxiety, and depression. People scoring low on Emotional Stability (high Neuroticism) are more likely to experience stress and nervousness, while people scoring high on Emotional Stability (low Neuroticism) tend to be calmer and self-confident.

The traditional approach to identify an individual's personality is through a survey. A well known test is the Big Five inventory questionnaire [37] which asks participants to evaluate on a 5 point scale how well their personality matches a series of descriptions for each of 5 personality traits. The Big 5 traits are then obtained by applying factor analyses to various lists of trait adjectives used in personality description questionnaires by using a statistical procedure called *Lexical Hypothesis* [3, 58]. The responses are then scored based on preassigned numbers to each reponses which are further added to obtain numerical scores in the range of 1-5 for each personality trait [29].

1.4 Problem Statement

In this study we investigate several approaches to infer the Big 5 personality scores from the user generated content on online social media like Facebook and YouTube vloggers. We use two kinds of dataset for this purpose, namely

- YouTube vloggers dataset which consists of transcripts of user's speech and several audio visual features from the video. Personality impression scores will be predicted from this dataset.

- myPersonality dataset which consists of users' online profile features and status updates text. Self reported personality scores of users will be predicted from this dataset.

Additionally, we examine the relationship between users' emotions and other characteristics on their Facebook profile like age, gender, time of posting, and personality by using an emotion detector algorithm we built for this purpose [21].

1.5 Contribution

We contribute to the research on personality recognition by exploring different approaches to personality recognition using a variety of feature set combinations by using single and multi target regression approaches. We show that using suitable techniques for feature selection, like correlation of a given feature with a personality trait, is vital to the quality of the prediction models generated. In addition to predicting each personality trait independently, we use multitarget regression models that will predict all 5 scores together by taking into account the correlations among the personality traits as an augmenting feature. Using this approach, we are able to outperform the average baseline model for all 5 personality traits. Infact our results reveal the state of the art average $RMSE$ of .76 in computational personality recognition from multimodal features for YouTube vloggers dataset [15]. We contribute to the emerging domain of personalised services by studying the relation between emotions expressed in approximately 1 million Facebook (FB) status updates and the users' age, gender and personality. Additionally, we investigate the relations between emotion expression and the time when the status updates were posted.

The remainder of the thesis is structured as follows: In Chapter 2 we describe the technical background of various regression models that we examine along with the feature selection tehniques that we use in our experiments. Next in Chapter 3, we describe the 2 datasets that we use in our experiments along with the features that we extract from them. In Chapter 4, we present the results of emotion analysis of Facebook posts and the relation between emotion and various Facebook profile features. In Chapter 5 we present the results of personality prediction on a YouTube vloggers and a myPersonality dataset using different techniques. Finally, we conclude with our overall findings in Chapter 6. Additionally, the correlations between the various features extracted from the 2 datasets and the personality scores are presented in the Appendix chapters.

Chapter 2

BACKGROUND

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." by Tom M. Mitchell [47]

Translating it to our work, experience E refers to the groud truth dataset comprising of labelled personality scores of users (explained in Chapter 3), task T refers to the task of inferring Big 5 personality traits of users (explained in Chapter 5) and performane P measured using $RMSE$ and R^2 (described in Section 2.2). Machine Learning concepts have wide application in different domains like search engines, computer vision, spam filtering etc. Based on the data available to the learning system, machine learning techniques can be classified as *Supervised* (data with known outcomes), *Unsupervised* (data without known outcomes) and *Re-inforced learning* (dynamic on-line performance like driving a car). Machine learning techniques can also be classified based on the task to be accomplished such as *Classification* (outcomes are class labels like yes/no), *Regression* (outcomes are continuous values like real numbers) etc. Depending on the goal of the target system that might use the personality traits, different types of models could be used to model Big 5 personality traits. It can be treated as a classification problem as in [22, 55] where some kind of thresholding is used to convert the scalar valued scores into classes. However, it could also be modelled as real valued scores using either regression [44, 5] or ranking algorithms [24]. In this work, we treat the this problem as a supervised regression task which is used to model the real valued scores as found in the personality ratings data. We will study the concept of regression and machine learning algorithms that perform regression in this chapter.

2.1 Regression Model for Personality Recognition

Regression is the task of predicting a continuous, real valued output from a set of predictors. In this work, we refer to univariate and multivariate regression as the model with one dependent variable (one outcome) and more than one dependent variables (several outcomes) respectively. Prediction tasks with multiple outcomes as in personality prediction (*Extraversion*, *Agreeableness*, *Conscientiousness*, *Emotional Stability/Neuroticm* and *Openness*) can be modelled as five univariate models one for each outcome which will be queried at prediction time separately. This is the univariate regression approach which is explained below using two algorithms namely *Linear Regression* and *Support Vector Machines.*

2.1.1 Univariate Regression Approach

1. *Linear Regression*

   Given a set of training data of the form $\{(x_1,y_1)\ldots(x_l,y_l)\} \subset \mathcal{X}\times\mathbb{R}$ where $\mathcal{X}$ represents input space (e.g $\mathcal{X} = \mathbb{R}^d$) in linear regression, the goal is to find a function $h_\theta(x)$ that has the least deviation from the actual targets of the training data y_i. A linear function can be described as -

   $$h_\theta(x) = \theta_0 + \theta_1 x^{(1)} + \theta_2 x^{(2)} + \ldots + \theta_d x^{(d)} \tag{2.1}$$

   where ($x^{(1)}$, $x^{(2)}$, ... $x^{(d)}$) is the input feature vector for the dataset $\mathcal{X}$,
   θ_i are the weights or parameters parameterizing the space of linear function from $\mathcal{X}$ to $\mathcal{Y}$.
   The Equation 2.1 can be simplified as -

   $$h(x) = \sum_{i=0}^{d} x_i\theta_i \tag{2.2}$$

   where the intercept term $x^{(0)}$ is 1. The value for θ are learnt by making the hypothesis function $h(x)$ close to the actual outcomes y^i for the training samples. We define a cost funtion, $J(\theta)$ that measures the closeness of $h(x_i)$ to the actual outcomes y_i.

   $$J(\theta) = \frac{1}{2}\sum_{i=1}^{l}(h_\theta(x_i) - y_i)^2 \tag{2.3}$$

This is referred to as the *least squares regression model.* The cost funcion is minimized in order to obtain optimum value for θ. This method begins with an initial value for the θ and progressively changing θ in order to minimize $J(\theta)$ until it converges to a minimum value. The update function if as follows -

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \tag{2.4}$$

where α is called the *learning rate.* This update is applied to all the θ_j where $j = 0$ ... d. This algorithm takes each step in the direction of lowering cost funtion J(θ). Solving the partial derivatives, we obtain the following update rule referred to as *Least Squares Update Rule* -

$$\theta_j = \theta_j + \alpha \sum_{i=1}^{l} (y_i - h_\theta(x_i)) x_i^j \quad \textit{for every } j \tag{2.5}$$

We can see that the magnitude of the update term is proportional to the error term (y_i - $h_\theta(x_i)$), as the deviation from the actual outcomes increases, the θ value increases too. This algorithm is called *Batch Gradient Descent* since the update looks at each of the training sample at each step of the descent. There are other alternatives for minimizing the cost function apart from this iterative algorithm, namely using derivatives of the input feature matrix.

2. *Support Vector Machines*

   Support Vector Machines (SVM) are a group of supervised machine learning techniques used for classification and regression tasks. SVM is based on statistical learning theory or VC theory [68] which charecterises learning machines to generalise to unseen data. The basic idea is as follows [65] -

   Given a set of training data of the form $\{(x_1,y_1) \ldots (x_l,y_l)\} \subset \mathcal{X} \times \mathbb{R}$ where $\mathcal{X}$ represents input space (e.g $\mathcal{X} = \mathbb{R}^d$) in ε-SV regression, the goal is to find a function $f(x)$ that has at most ε deviation from the actual targets of the training data y_i and is as flat as possible. Errors are permitted until they are at most ε beyond which they are not permitted. A linear function can be described as -

$$f(x) = \langle w, x \rangle + b \tag{2.6}$$

with $w \in \mathcal{X}, b \in \mathbb{R}, \langle ., . \rangle$ denotes dot product in $\mathcal{X}$. The regression function of the SVM will use a penalty only if the predicted value $f(x)$ is more than ε distance away from the actual value y_i. Flatness in this case would mean small values for w which is obtained by convex optimization. In this case, we assume that such a function f is feasible which approximates all the data $\langle x_i, y_i \rangle$ within ε precision. If not, slack variables are introduced to cope with the infeasible convex optimization problem. Hence the convex optimization problem involves minimization of -

$$\frac{1}{2}||w||^2 + C\sum_{i=1}^{l}(\varepsilon_i + \varepsilon_i^*) \tag{2.7}$$

where ε_i and ε_i^* are slack variables and $C > 0$ is the constant which determines the tradeoff between the flatness of f and amount upto which deviations larger than ε can be accomodated. As shown in Figure 2.1, points lying outside the shaded region incur cost as the deviations are penalised using the slack variables. The equation 2.7 can be solved by dualization using Lagrange multipliers. We arrive at our final function $f(x)$ as described below -

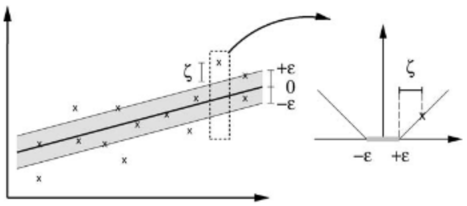


Figure 2.1: Soft Margin loss setting in linear SVM [65]

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\langle x_i, x \rangle + b \tag{2.8}$$

where α_i and α_i^* are Lagrange multipliers. We can see that w in Equation 2.8 can be specified by the linear combination of training examples x_i. It follows that for all samples inside the ε tube which is the shaded region in the Figure 2.1, Lagrange multipliers α_i and α_i^* are zero whereas samples outside the ε tube have non zero co-efficients. Hence in order to define w we only need non vanishing co-efficients which come from samples outside the ε tube, which are called the *Support Vectors*. In case of non-linear separation between the sample data, in order for SVM to be able to find a suitable hyperplane, we need to project this sample into higher dimensions for SVM to be able to find a hyperplance. Kernels are used in SVMs for such non-linear operation in which the input samples are mapped using a mapping function $\Phi : \mathcal{X} \rightarrow \mathcal{F}$. Then standard SV regresion is applied on the higher dimensional feature space. The mapping function is called the kernel function. Some common kernel functions include polynomial, Gaussian radial basis, hyperbolic etc.

2.1.2 Multivariate Regression Approach

Multivariate regression, which is interchangeably called multi-output or multi-target regresison, aims at predicting multiple real valued outputs simultaneously instead of independent ones. The results in Table 2.1 and Table 2.2 indicate, there is a clear correlation among different personality trait impression scores in the YouTube and myPersonality datasets. For a more detailed explanation about correlation analysis, we refer to Section 2.3.1. This makes personality score prediction a good candidate for multivariate regression, where the dependencies between the target variables are taken into account to make a combined prediction. A complete description of the 2 datasets is provided in Chapter 3. Formally, multivariate regression addresses the following problem:

Let F be the vector (feature space or input space) including m features, $f_1, f_2, ..., f_m$, and T be the target vector (output space) including n target variables $t_1, t_2, ..., t_n$. The goal of a multivariate regression algorithm is to learn a model $M : F \rightarrow T$ that minimizes the prediction error over a test test. Using this formulation, the 6 multivariate regression algorithms that we use in this paper are [74]:

1. **Single Target (ST)**: In ST, for each target variable t_i, a single model is trained based on the input vector F (i.e., $F \rightarrow t_i$). The results of the multi-target model are comprised of all n single target ones.

2. **Multi-Target Stacking (MTS)**: MTS consists of two steps. In the first step, n single-target models are used as in ST, however, MTS includes an additional step

Table 2.1: Pearson product-moment correlation results among personality scores on 5 traits: *Extraversion (Extr)*, *Agreeableness (Agr)*, *Conscientiousness (Cons)*, *Emotional Stability (EmoStab)*, *Openness (Open)* on YouTube vloggers dataset. Significant correlations ($p < .05$) among the personality scores are indicated in bold.

	Extr	Agr	Cons	EmoStab	Open
Extr	1.00				
Agr	.02	1.00			
Cons	-.03	**.38**	1.00		
EmoStab	.06	**.69**	**.54**	1.00	
Open	**.56**	**.29**	**.26**	**.30**	1.00

Table 2.2: Pearson product-moment correlation results among personality scores on 5 traits: *Extraversion (Extr)*, *Agreeableness (Agr)*, *Conscientiousness (Cons)*, *Neuroticm (Neu)*, *Openness (Open)* on myPersonality dataset. Significant correlations ($p < .05$) among the personality scores are indicated in bold.

	Extr	Agr	Cons	Neu	Open
Extr	1.00				
Agr	**.16**	1.00			
Cons	**.18**	**.17**	1.00		
Neu	**-.33**	**-.34**	**-.29**	1.00	
Open	**.14**	**.04**	**.02**	**.30**	1.00

where the input space for each target variable is expanded by the predicted results of the other target variables ($n - 1$ predicted values) from step one. Let $t'_1, t'_2, ..., t'_n$ be the prediction results from the first step, then for example the input space for t_1 in step two is $[f_1, f_2, ..., f_m, t'_2, t'_3, ..., t'_n]$.

3. **Multi-Target Stacking Corrected (MTSC)**: In MTSC, an internal cross validation sampling technique is used to avoid over-estimation of the training set. In MTSC, by using k-fold sampling, the prediction results of $\frac{k-1}{k}\%$ of the whole training set are used to expand the input space in the second step as in MTS.

4. **Ensemble of Regressor Chains (ERC)**: The idea behind ERC is chaining single-target regression models. By choosing an order for the target variables (e.g., $O = (t_1, t_2, ..., t_n)$), the learning model for each target variable t_j relies on the prediction results of all target variables t_i which appear before t_j in the list. For the first target variable, a single-target regression model as in ST predicts the value, then the input space for the next target variable is extended with the prediction results of the previous one and so on. Since in this model the order of the chosen chain affects the results, the average prediction result of r different chains (typically $r = 10$) for each target variable is used as the final prediction result.

5. **Ensemble of Regressor Chains Corrected (ERCC)**: The difference between ERC and ERCC is similar to that between MTS and MTSC, i.e. the use of k-fold sampling to increase the reliability of the predictions based on the training set.

6. **Multi-objective random forest (MORF)**: MORF is based on ensembles of multi-objective decision trees. We refer to [41] for further explanation.

Note that ST does not leverage the prediction result for one personality trait to make a prediction for another, while all other algorithms (MTS, MTSC, ERC, ERCC and MORF) do in one way or another. For the results in Chapter 5 we used the implementation of these algorithms in Mulan.[1] All algorithms except MORF use Weka decision trees as a base learner. For further information we refer to [74].

2.2 Evaluation

2.2.1 Root Mean Squared Error

It is the measure of the difference between the predicted values by a model and the observed values. It is computed by taking the square root of the average of the square of the differences. It is a measure of the closeness of the fitted line to the data points. Since the errors

[1] http://mulan.sourceforge.net/

are squared, this measure is sensitive to large errors. RMSE ranges from 0 to ∞ where lower values signify better models. RMSE can be described by the following formula -

$$\text{RMSE}= \sqrt{\frac{\sum_{t=1}^{n}(y_{obs}^{t}-y_{pred}^{t})^2}{n}}(2.9)$$

where y_{obs} and y_{pred} are the observed and predicted scores for instance t where (t = 1 ... n) and n is the sample size.

2.2.2 Co-efficient of Determination

Co-efficient of determination (R^2) is the ratio of the model's absolute error and the baseline mean predicted scores. It is expressed as -

$$\text{R}^2 = 100 \times (1 - \frac{\sum_{t=1}^{n}(y_{obs}^{t}-y_{pred}^{t})^2}{\sum_{t=1}^{n}(y_{obs}^{t}-\hat{y}_{obs}^{t})^2})(2.10)$$

where y_{obs} and $\hat{y}_{obs}$ are the observed scores and its mean respectively and y_{pred} are the predicted scores by the model. It measures the relative improvement of the Mean Squared error using the model compared to the baseline. Positive values indicates that the model outperformed the baseline whereas negative values indicates that model did not outperform the baseline.

2.2.3 k-fold Cross Validation

Cross validation is a model validation technique to assess how well the learning model generalises to an unseen dataset. For this purpose, we set aside a part of the training dataset before the training phase. After training the model, it is then tested on the reserved dataset to test the performance of the learned model. With this general idea, there are several variations of cross validation like holdout method, leave one out, k-fold cross validation etc. The latter technique is an improvement over the holdout method, in which the dataset is randomly split into training and test set with the training set used for learning the model and the test set used in evaluating the model. K-fold cross validation is an extension of the holdout method in which the dataset is split into k folds. Each fold gets to be tested once and the training occurs using the other $(k-1)$ folds. Hence each datapoint gets to be tested once and trained $(k-1)$ times resulting in better prediction error measures. The average error across all k folds is computed.

2.3 Feature Selection Methodology

Feature subset selection is the process of identifying relevant features and removing irrelevant and redundant features during the training of the model. It has been shown that feature subset selection enhances the performance of learning algorithm by reducing the hypothesis search space and/or reducing the storage or processing requirement [31]. Hence, our goal would be to identify features that are *most predictive and relevant* to the target variable. Feature selection algorithms perform an exhaustive search through the feature subset and hence the following four different issues need to be addressed [11] by such a feature selection algorithm:

1. *Starting Point*

   This defines the direction of search. For instance, *forward search* begins with an empty feature subset with successive feature addition at each step whereas *backward search* begins with the full feature space with successive deletions at each step.

2. $Organization\ of\ search$

   Since an exhaustive search through the entire feature subspace would involve an exponential number of subsets to evaluate, a predetermined method to navigate through the feature subspace is essential. *Stepwise selection of elimination* is an example, in which a decision to add or remove the feature is made at each step or decision point.

3. $Evaluate\ feature\ subset$

   Strategies to evaluate feature subsets lead to two different approaches, namely *filter and wrapper* methods. Feature subsets are evaluated independent of the learning algorithm in filter methods. In contrast, wrapper methods use the error measures generated by training the learning algorithm on the dataset based on the candidate feature sets to evaluate the given feature subsets.

4. *Stopping Criteria*

   The decision to stop the search involves different strategies for filter and wrapper methods. In case of wrapper methods, the search can be stopped when adding or deleting features at each step does not change the error measure of the learning algorithm. Alternatively, for the filter methods, the search can be stopped by ordering the feature subsets based on a computed score and selecting features using a threshold on the computed score.

For our work, we set values for the above 4 parameters as follows -

- $Starting\ Point$ - Empty feature set.
- $Organization\ of\ search$ - Stepwise addition of feature subsets until all feature subsets are considered.
- $Evaluate\ feature\ subset$ - $RMSE$(refer Chapter 2.2) is used to evaluate each feature subset using SVM learning algorithm(refer Chapter 2.1).
- $Stopping\ Criteria$ - We stop forward search after we examine all feature subsets.

2.3.1 Correlation Analysis

Correlation analysis measures the strength of the relation that exists between two variables. In our work, we use Pearson and Spearman correlation measures. Pearson correlation is the most common way to measure the strength of the linear relationship between variables. The most common formula for computing Pearson product-moment correlation co-efficient (r) is -

$$r = \frac{\sum_{x=1}^{n} xy}{\sqrt{(\sum_{x=1}^{n} x^2) \cdot (\sum_{x=1}^{n} y^2)}} \tag{2.11}$$

where (x_1,x_2 $\cdots$ x_i) and (y_1,y_2 $\cdots$ y_i) are the 2 pairs of observations, x = x_i - $\bar{x}$ and y = y_i - $\bar{y}$.

The sign and value of the correlation co-efficient suggests direction and strength of the relation between the two variables. The greater the magnitude of the correlation, the stronger the linear relationship. Positive correlation indicates that both the variables increase or decrease unidirectionally. A negative correlation suggests an inverse relation between the two, which means that as one variable increases the other decreases and vice versa. Since the Pearson correlation only measures a linear relationship, even if the correlation is 0, it only signifies that there is no linear relationship between them. Quadratic or curvilinear relationships could still exist between the two variables.

Spearman rank correlation co-efficient is the non-parametric version of the Pearson correlation and measures the association between ranked variables. In this work, we also use Spearman since the assumptions of Pearson correlation were violated for one of our datasets. Spearman is used to measure the monotonic relationship between two variables by computing its ranks based on the value of the variable. The formula for computing Spearman rank-order correlation(ρ) for tied ranks is as follows -

$$\rho = \frac{\sum_{x=1}^{n} xy}{\sqrt{(\sum_{x=1}^{n} x^2) \cdot (\sum_{x=1}^{n} y^2)}} \tag{2.12}$$

where (x_1,x_2 $\cdots$ x_i) and (y_1,y_2 $\cdots$ y_i) are the 2 pairs of observations, x = x_i - $\bar{x}$ and y = y_i - $\bar{y}$.

It can be seen that this is similar formula as Pearson. When there are ties ranks, the Spearman correlation is the Pearson correlation co-efficient between the ranks. The formula for tied ranks are -

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(d_i^2)}{n(n^2-1)} \tag{2.13}$$

where d_i is the difference in paired ranks between 2 pairs of observations and n = number of observations.

In either methods for computing correlation, we use significance tests to find significantly correlated features by setting α to .05. We measure significance by setting the null and alternate hypothesis as follows:

H_0 (Null Hypothesis) : There is no association between the two vaiables.

H_1 (Alternate Hypothesis) : There is an association between the two variables.

If the obtained p value is less than the predetermined α value, we will reject the null hypothesis and accept the alternate hypothesis. Such a relation is statistically significant.

2.3.2 Forward Seach Feature Selection using Correlation

Research in feature selection literature suggests that irrelevant and redundant features need to be eliminated from the feature space for the learning algorithm to perform well. To this end, we follow this idea -

A good feature subset contains features highly corelated (predictive of) with the target variable, yet uncorrelated (not predictive of) with each other. [31]

While relief [40], Minimum Description Length, symmetric uncertainty are used as feature weighting measures in the above study, we use Pearson and Spearman correlation based heuristics to evaluate feature subsets. In particular we use wrapper strategy to evaluate the relative merit of correlated features over using all features. The features subsets (correlated and all) are used to train two different learning algorithms - Linear Regression and Support Vector Regressors. In this way, we combine the filter method (selecting features with significant correlation with the target variable) and wrapper method (evaluate the performance of the feature subset on a given ML algorithm). The feature subset evaluation function is the Root Mean Squared Error(RMSE) given here -

$$\text{RMSE}= \sqrt{\frac{\sum_{t=1}^{n}(y_{obs}^{t}-y_{pred}^{t})^2}{n}} (2.14)$$

where y_{obs} and y_{pred} are the observed and predicted scores for instance t where ($t = 1 \ldots n$) and n is the sample size. The individual features are grouped as feature subsets based on the category of the features, thus feature subsets are groups of at least 2 individual features. Our starting point of forward search is the feature subset with the lowest RMSE value. We search through the feature space by progressively adding feature subsets that will improve the RMSE value or leave it unchanged. We continue this process until we examine all the feature subsets. Hence at the end of forward search, either an feature subset with improved or unchanged $RMSE$ is obtained. We evaluate the performance of the feature subset obtained by forward search on the test dataset to obtain our final $RMSE$ value.

Chapter 3

DATASETS DESCRIPTION AND PREPARATION

All the results presented in this thesis are based on 2 different datasets. We will describe each dataset in full detail in this chapter.

3.1 myPersonality Dataset

Our results for personality prediction (refer Chapter 5.3.2) and experiments on emotion analysis of users' status updates (refer Chapter 4) are based on the data from the myPersonality project [43]. myPersonality was a popular Facebook application introduced in 2007 in which users took a standard Big Five Factor Model psychometric questionnaire [30] and gave consent to record their responses and Facebook profile. The survey takers were highly motivated to answer honestly and carefully since they received feedback on their personality results for their participation. This data consists of information about users' demographics (e.g., age and gender), friendship links, Facebook activities (e.g., number of group affiliations, number of status posts, page likes, education and work history), status updates (time and text of the post) and Big Five Personality Scores (on a scale from 1-5). The Table 3.1 represents the count of users in the entire dataset in different tables.

However, not all of this information is available for all users. Hence, from this data, we make 2 different datasets for each of the 2 experiments mentioned above. Additionally the dataset described below in Section 3.1.1 is a subset of the dataset used for personality prediction described in Section 3.1.2. The difference between the 2 datasets is that we treat each status update posted by an user individually for emotion analysis whereas we combine all the status updates posted by an user and treat them as one document for the personality prediction problem.

Category	Number of users
Facebook Activity	1.7mn
Big5	3.1mn
Demographic	4.3mn
Status updates	154k (220mn posts)

Table 3.1: myPersonality data statistics

3.1.1 Dataset for Emotion Analysis of Facebook Posts

For this experiment (refer Chapter 4), we make a dataset of 5,865 users for which we have information about their age, gender, personality scores and at least one status update. Table 3.2 provides details about this dataset's characteristics. The dataset contains personality scores ranging from 1 to 5 for each user and each personality trait. To facilitate further analysis, for each personality trait, we split the set of users into those that clearly exhibit the trait and those who do not. Towards this end we use the same thresholds that were used in the WCPR13 data set[1]. The score threshold and the number of users for each personality trait is presented in Table 3.2. For instance, in the remainder of the Chapter 4, we call a user an extrovert if his Extroversion score is at least 3.60; there are 2,971 such users in our data set. Note that such a binary split of users along the 5 personality dimensions is a fairly crude approach, and that a more fine grained study that considers the sliding scale from Introversion to Extroversion could provide further insights.

3.1.2 Dataset for Personality Prediction

For this experiment (refer Chapter 5), we make a dataset of 38,106 users for which we have information about their demographic profile (age and gender), Facebook activity (count of status posts, network size, groups, likes, diads and education) and their status updates. Additionally, we select only those users, for which at least one status update is available and

[1] http://mypersonality.org/wiki/doku.php?id=wcpr13

Table 3.2: (Table on the left) Characteristics of female and male users in the dataset. The entire dataset contains 969,035 status updates written by 5,865 users. (Table on the right) Score threshold and number of users for each personality trait. Note that the same user can exhibit more than one personality trait at once.

	Female	Male
# users	3,446	2,419
Average age	26	25
# posts	625,921	343,114
Avg # posts/user	182	142
Min # posts/user	1	1
Max # posts/user	2,428	1,453

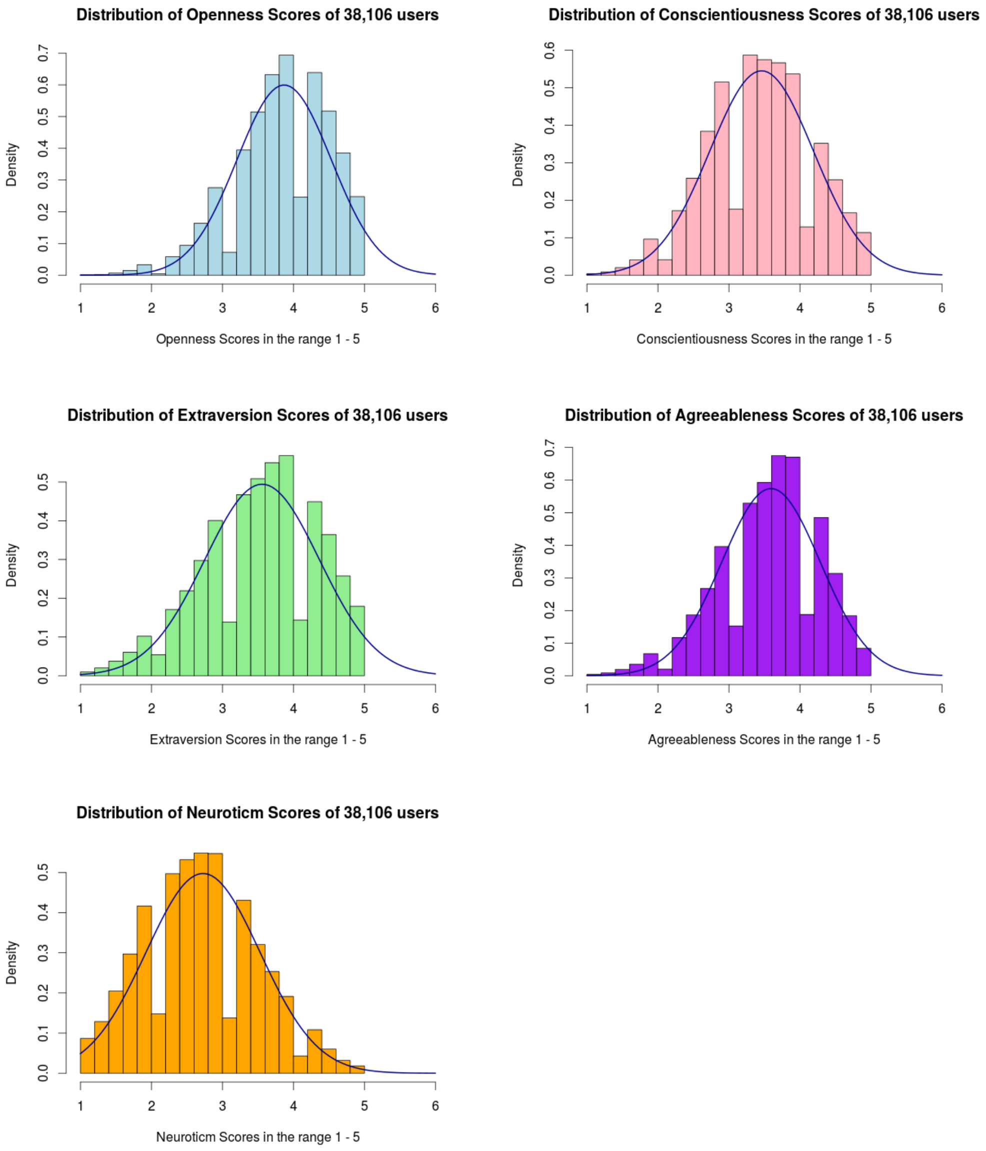
Personality	Threshold	# of users
Extroversion	3.60	2,971
Openness	3.80	3,284
Agreeableness	3.55	3,110
Conscientiousness	3.50	3,071
Neuroticism	2.80	2,631

English is chosen as the language version of their Facebook profile. Our dataset consists of a total of 6,918,789 status updates of 38,106 users. Since our goal is to infer the Big 5 personality scores for a given user, we identify a user with their set of available status updates (treated together as one text per user when extracting linguistic features), their demographic features and Facebook activities. The Big 5 personality scores for each of the 5 traits are available for each user in the range of 1-5. The figures (refer Figure 3.1) show the distribution of personality scores among the 38,106 users which follows a normal distribution curve.

Table 3.3 provides details about this dataset's characteristics and personality scores distribution.

3.2 myPersonality Dataset - Features Used

We extracted a wide variety of linguistic and emotional features from the status updates of the users. The underlying rationale for including linguistic and emotional features is that people with different personality traits will express themselves differently and, hence, will use different words (phrases) and emotions (anger, joy) when expressing themselves. A relation between emotions and personality traits has been observed in past research as well



Distribution of Openness Scores of 38,106 users

Distribution of Conscientiousness Scores of 38,106 users

Distribution of Extraversion Scores of 38,106 users

Distribution of Agreeableness Scores of 38,106 users

Distribution of Neuroticm Scores of 38,106 users

Figure 3.1: Distribution of the Big 5 Scores - Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticm scores among the 38,106 users.

[21]. Overall, we use the following features[2].

[2] See the appendix B for a full list of features and how they correlate with the 5 personality traits. All features except for the gender feature are numerical.

Table 3.3: (Table on the left) Characteristics of 38,106 users in the dataset. (Table on the right) Mean and Standard Deviation of Big 5 personality scores of 38,106 users.

	Female	Male
# Users	22,358	15,748
Average age	26	25
Avg Network size/user	305	302
Avg # Likes/user	196	153
Avg # Diads/user	226	208
Avg # Education/user	2	2
Avg # Status Updates/user	202	153
Avg # Groups/user	32	33

Personality	Average Score	Std Dev
Extroversion	3.56	.81
Openness	3.87	.67
Agreeableness	3.6	.69
Conscientiousness	3.46	.73
Neuroticism	2.73	.8

1. **Demographic Features:** 3 features related to the demographic profile of the user: (1) Age and (2) Gender of the user.

2. **Facebook activity:** Features related to their activity on the website including (1) Count of likes, (2) Count of status updates posted by the user, (3) Count of education, (4) Count of diads from the friendship diads table of the user and (5) Count of group memberships for the user (6) Network size or number of friends of the user.

3. **Linguistic Features:**

    - **LIWC:** 81 features extracted using the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker and King 1999), consisting of features related to (1) standard counts (e.g., word count), (2) psychological processes (e.g., the number of anger words such as hate, annoyed ... in the text), (3) relativity (e.g., the number of verbs in the future tense), (4) personal concerns (e.g., the number of words that refer to occupation such as job, majors ...), (5) linguistic dimensions (e.g., the number of swear words). For a complete overview, we refer to (Tausczik and Pennebaker 2010).

- **MRC**: MRC is a psycholinguictic database[3] which contains psychological and distributional information about words. The MRC database contains 150,837 entries with information about 26 properties (e.g., the number of syllables in the word, the number of letters, etc.), although not all properties are available for every word. Using MRC we generated **14 features** for every status update by adding the MRC scores for each word in the combined status post for each user.
- **SentiStrength:** SentiStrength[4] assigns to each text a positive and negative sentiment score on a scale of 1 (no sentiment) to 5 (very strong sentiment). Posts may be simultaneously positive and negative to varying degree. We used SentiStrength to compute 2 sentiment scores (**2 features**) for every combined status update for each user.
- **SPLICE:** We used SPLICE[5] (Structured Programming for Linguistic Cue Extraction) to extract **71 linguistic features**, including cues that relate to the positive or negative self evaluation of the user (e.g., *I'm able*, *don't know*), complexity scores (e.g., average word length, average sentence length ...), scores based on reading (e.g., FOX, LIX, DALE ...) style of the user.

3.3 YouTube vloggers dataset

The YouTube personality dataset[6] consists of a collection of *audio-video features*, *speech transcripts*, *gender*, and *personality impression scores* for a set of 404 YouTube vloggers. We used the split of the data into 348 training and 56 test instances that was suggested by the organizers of WCPR2014 [7]. The vloggers explicitly show themselves in front of a webcam, talking about a variety of topics including personal issues, politics, movies, books, etc. Figure 3.2 shows an excerpt from the transcript of a vlogger. The *audio-video features* were automatically extracted from the conversational excerpts of vlogs and

[3] http://www.psych.rl.ac.uk/User_Manual_v1_0.html

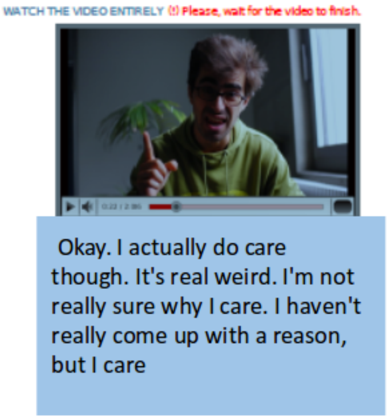
[4] http://sentistrength.wlv.ac.uk

[5] http://splice.cmi.arizona.edu

[6] https://www.idiap.ch/dataset/youtube-personality

[7] https://sites.google.com/site/wcprst/home/wcpr14

aggregated at the video level. The *speech transcripts* correspond to the full video duration. The transcripts are provided in raw text and contain a total of approximately 10K unique words and approx. 240K word tokens. Finally, the *personality impressions* consist of Big Five personality scores that were collected using Amazon Mechanical Turk and the Ten-Item Personality Inventory (TIPI). MTurk annotators watched one-minute slices of each vlog, and rated impressions using a personality questionnaire. The Big 5 personality impression scores are available for each user in the range of 1.5-6.5 over all the 5 traits. Table 3.4 provides details about this dataset's characteristics and personality scores distribution. The figures (refer Figure 3.3) show the distribution of personality scores among the 404 vloggers which follows a normal distribution curve.



WATCH THE VIDEO ENTIRELY (!) Please, wait for the video to finish.

Okay. I actually do care though. It's real weird. I'm not really sure why I care. I haven't really come up with a reason, but I care

Figure 3.2: An example of an excerpt from a vlogger transcript

3.4 YouTube vloggers Dataset - Features Used

In addition to the given audio/video and age features, we extracted a wide variety of linguistic and emotional features from the vlogger transcript. We used Spearman's rank correlation coefficient to assess the strength of the relationship between the different features described below and the 5 perceived personality traits (see the appendix). Given the highly skewed distribution of some the features, we decided to report all the correlations using Spearman's coefficient, which is better suited for non-normal data. Our results indicate a strong

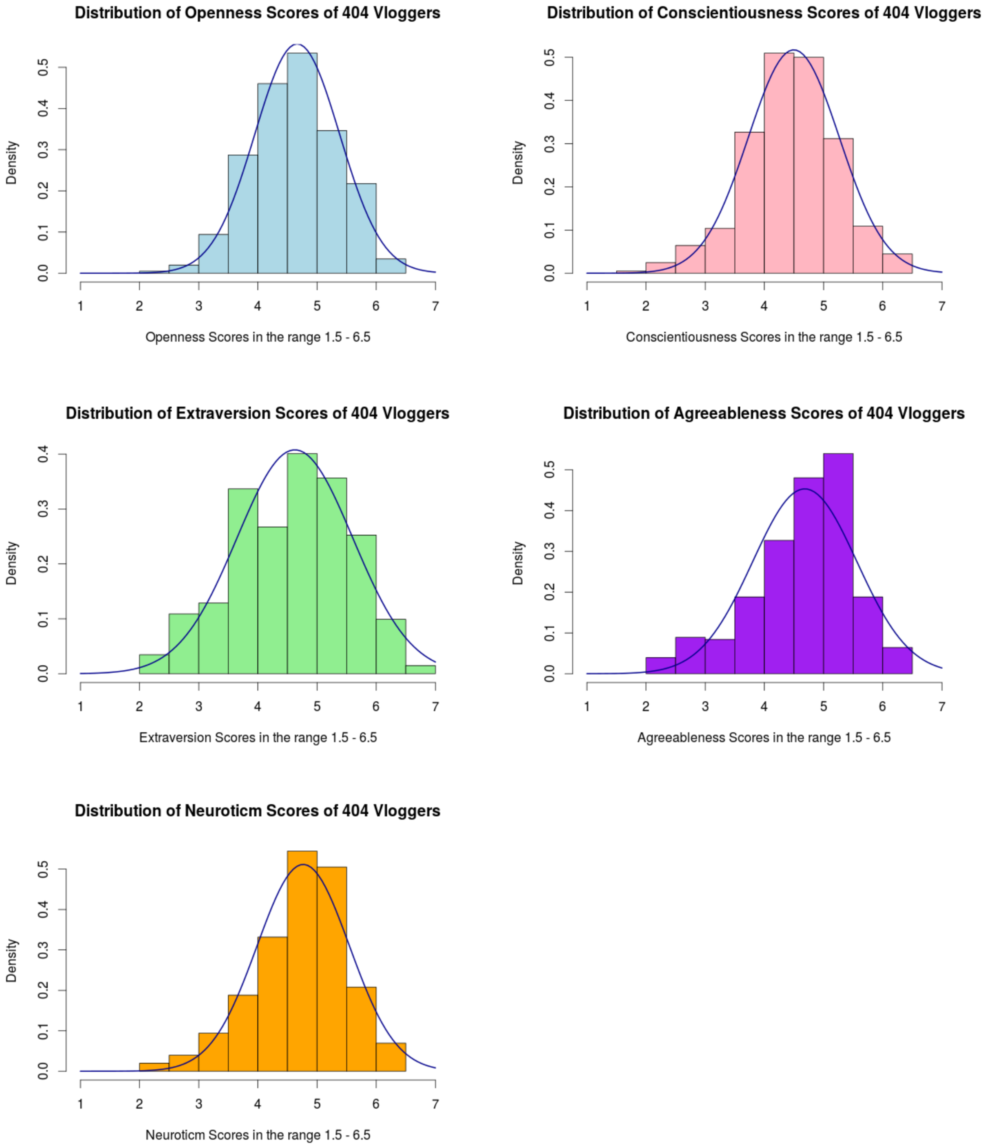


Figure 3.3: Distribution of the perceived Big 5 Scores - Openness, Conscientiousness, Extraversion, Agreeableness, Emotional Stability scores among the 404 vloggers.

relationship between many linguistic and emotional features and personality impressions. Motivated by previous research, and the observed correlation between features and personality impressions, we decided to include these features in our regression models. Some of our feature sets have some semantic overlap (e.g. NRC and SentiStrength, and LIWC and

Table 3.4: (Table on the left) Characteristics of 404 users in the YouTube vloggers dataset. (Table on the right) Mean and Standard Deviation of perceived Big 5 personality scores of the users.

Dataset	Charecteristics
# users	# Female - 210
	# Male - 196
# Audio/Video features	# Audio - 21
	# Video - 4
Transcripts	10K unique words
	250k word tokens
	Avg 595 words/transcript

Personality	Average Score	Std Dev
Extroversion	4.62	.98
Openness	4.66	.72
Agreeableness	4.68	.88
Conscientiousness	4.5	.77
Emotional Stability	4.77	.8

SPLICE). The use of feature selection methods to be more selective in the choice of features is an interesting direction for further research.Overall, we used the following features[8]

1. **Gender:** We used a binary *gender* feature to identify male and female. Overall, the data is balanced in terms of gender distribution and includes 210 females (52%) and 194 males (48%).

2. **Audio-Video:** We used all 25 audio-video features that are provided with the vlog dataset [9]. These include speaking activity and prosody cues such as speaking time and pitch, as well as video features such as the number of look turns and camera proximity.

3. **LIWC**: From the transcripts we extracted **81 features** using the Linguistic Inquiry and Word Count (LIWC) tool [60], including features related to standard counts (e.g., word count), psychological processes (e.g., the number of anger words such as *hate* and *annoyed* in the transcript), relativity (e.g., the number of verbs in the future

[8] See the appendix for a full list of features and how they correlate with the 5 perceived personality traits. All features except for the gender feature are numerical.

tense), personal concerns (e.g., the number of words that refer to occupation such as *job* and *majors*), and linguistic dimensions (e.g., the number of swear words).

4. **NRC:** NRC is a lexicon that contains more than 14,000 distinct English words annotated with 8 emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust), and 2 sentiments (negative, positive) [50]. For each transcript we counted the number of words in each of the 8 emotion and sentiment categories, resulting in **10 features** per transcript.

5. **MRC**: MRC is a psycholinguictic database[9] which contains psychological and distributional information about words. The MRC database contains 150,837 entries with information about 26 properties (e.g., the number of syllables in the word, the number of letters, etc.), although not all properties are available for every word. Using MRC we generated **14 features** for every transcript by adding the MRC-scores for each word in the transcript.

6. **SentiStrength:** SentiStrength[10] assigns to each text a positive, negative and neutral sentiment score on a scale of 1 (no sentiment) to 5 (very strong sentiment). Texts may be simultaneously positive, negative and neutral. We used SentiStrength to compute 3 sentiment scores (**3 features**) for every transcript.

7. **SPLICE:** We used SPLICE[11] (Structured Programming for Linguistic Cue Extraction) to extract **74 linguistic features**, including cues that relate to the positive or negative self evaluation of the speaker (e.g., *I'm able*, *don't know*), complexity and readability scores.

The underlying rationale for including linguistic and emotional features is that people with different personality traits will express themselves differently and, hence, will use

[9] http://www.psych.rl.ac.uk/User_Manual_v1_0.html

[10] http://sentistrength.wlv.ac.uk

[11] http://splice.cmi.arizona.edu

different words (phrases) and emotions (anger, joy) when expressing themselves. A relation between emotions and personality traits has been observed in past research as well [21]. We used Spearman's rank correlation coefficient to assess the strength of the relationship between the different features described above and the 5 perceived personality traits (see the appendix). Given the highly skewed distribution of some the features, we decided to report all the correlations using Spearman's coefficient, which is better suited for non-normal data. Our results indicate a strong relationship between many linguistic and emotional features and personality impressions. Motivated by previous research, and the observed correlation between features and personality impressions, we decided to include these features in our regression models. Some of our feature sets have some semantic overlap (e.g. NRC and SentiStrength, and LIWC and SPLICE). The use of feature selection methods to be more selective in the choice of features is an interesting direction for further research.

Chapter 4

EMOTIONS AND PERSONALITY IN FACEBOOK

4.1 Introduction and Motivation

Personality can affect the decision making process and has been shown to be relevant in the selection of music, movies, TV programs and books. It has been shown that personality affects preference for websites [42], language used in online social media [66], choice of Facebook Likes [43], music taste [64], and content such as movies, TV shows, and books [14]. In addition, it has been shown that users' emotions can also be used to detect users' taste at any moment, e.g., sad users are more likely to prefer action movies to watch [34]. Going yet one step further, personalized services can even have an impact on users' feelings. A nice example of this is that watching movies can change users' emotion, e.g., people feel joy when watching comedies or sadness when watching a late night romantic movie [34].

An interesting difference between personality and emotion is that personality is a stable characteristic and emotions are of short term duration. Emotion can be a momental feeling with respect to an object, person, event, or situation. As a consequence, people express a variety of different emotions over a period of time which is not the case for users' personality.

In this study, we detect emotions from users' status updates using the NRC word-emotion lexicon [52], and determine the relation between users' feelings and their demographics (age and gender) and personality. We also extract time features from the time stamp of the status updates to find the relation between users' emotions and time. Little work has been done that examines the relation between a user's emotions and other characteristics in social media. In [12] the authors extract emotions from Twitter posts and find correlations with major events in politics and popular culture during a specific time frame, but they focus on the public emotion as a whole and not on feelings or other characteristics of individual users. To the best of our knowledge, no work has been done to find the relations between different emotions and personality with respect to time factors. In [56] the authors study the relation between emotions and time, however their work is based on a questionnaire and not based on social media content. In [51], the authors use SVM classifiers to predict personality using emotion expression in text. For their experiments, they use essays from psychology students, while in this work we focus on emotion expression in Facebook status updates and its relation with users' personality.

4.2 Emotion Detection

To detect users' emotions from their status updates, we use the NRC hash-tag emotion lexicon [52]. This lexicon contains a 10-dimensional binary emotion vector for 14,177 English words. The 10 dimensions or emotion categories are: *positive*, *negative*, *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, and *trust*. In the NRC lexicon, *positive* and *negative* are actually referred to as sentiments instead of emotions, but in our study we use the terms emotions and feelings loosely and interchangeably to refer to all 10 categories of the NRC lexicon.

A word can convey several emotions at the same time. For instance, according to the NRC lexicon, "happy" represents positive, anticipation, joy, and trust emotions, while "birthday" represents positive, anticipation, joy, and surprise emotions. In the remainder of this chapter, we say that a status update conveys an emotion if it contains at least one term from the lexicon that is associated with that emotion. For example, the status update "thanks to everyone who wished me a happy birthday today" conveys positive, anticipation, joy, trust, and surprise emotions because of the presence of the words "happy" and "birthday". The other words in this particular status update do not convey any emotion according to the NRC lexicon. Figure 4.1 presents the frequency of emotions in the posts in our data set described in Chapter 3.1.1. Almost 60% of the status updates express at least one kind of emotion, and the positive emotion is clearly the most prominent one. For completeness, we point out that to detect emotions we only scan the status updates for exact occurrences of words from the NRC lexicon. We use a bag of words approach and do not consider any misspellings (e.g., hapy or haaaappy), negation (e.g., not good), strength of the emotions using adjectives or adverbs (e.g., very happy vs. happy) or combined words (e.g., long-awaited vs. long awaited). Moreover, any emotions expressed with words that

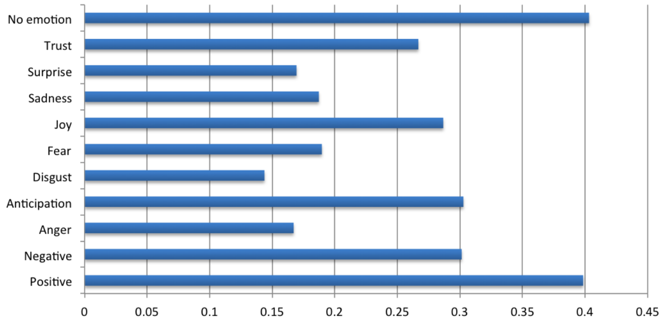


Figure 4.1: Emotion frequency in Facebook status updates. Almost 60% of the status updates express at least one kind of emotion from the NRC lexicon, and many posts convey more than one emotion.

are not present in the NRC lexicon will remain undetected.

4.3 Results Discussion

In the remainder of this chapter, let S denote the set of the 969,035 status updates in our study. Furthermore, for each of the 10 emotions 1:positive, 2:negative, 3:anger, 4:anticipation, 5:disgust, 6:fear, 7:joy, 8:sadness, 9:surprise, and 10:trust, let S_i, $i = 1, \ldots, 10$, be the set of status updates that contain at least one word associated with the respective emotion according to the NRC lexicon. As explained in Section 4.2, the sets S_1, S_2, …, S_{10} are not necessarily disjoint. In addition, we also introduce S_0 as the set of status updates that do not contain a term from the NRC lexicon, i.e. S_0 is the set of status updates that do not convey any emotion. It holds that $S = S_0 \cup S_1 \cup S_2 \cup \ldots \cup S_{10}$.

4.3.1 Emotion and Gender

Let S_f denote the set of status updates written by female authors and S_m the set of status updates by male authors. From Table 3.2 we know that women post more frequently than men. The probability that a status update is written by a woman is $P(S_f) \approx 0.65$ while the probability that it is written by a man is $P(S_m) \approx 0.35$. To determine the probability that a post conveys a particular emotion, given that it is written by a man or a woman, we calculate

$$P(S_i|S_m) = \frac{|S_i \cap S_m|}{|S_m|} \text{ and } P(S_i|S_f) = \frac{|S_i \cap S_f|}{|S_f|}$$

for $i = 0, 1, \ldots, 10$. The results are visualized in Figure 4.2. Although the differences between both genders are small, we do observe that female users in general express more emotions in their posts. In particular, women are more likely than men to post about positive feelings, joy and anticipation, while men are more likely than women to post status updates that convey anger or no emotion at all.

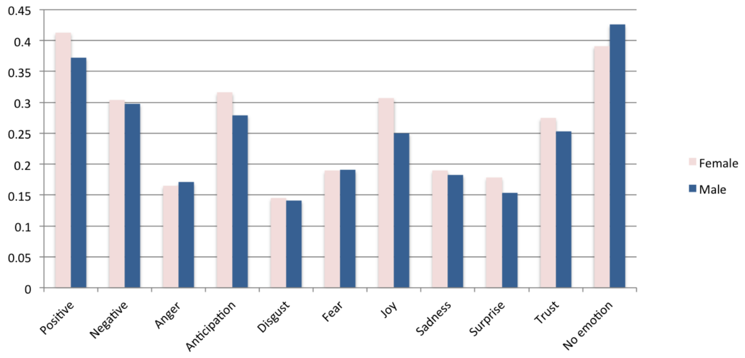


Figure 4.2: Probability of occurrences of emotions in status updates from female and male users.

4.3.2 Emotion and Age

To assess the relation between different age groups and their emotion expression in Facebook, we use five age groups: users younger than 21, users between 21 and 30, users between 31 and 40, users between 41 and 50, and users older than 51. The average age of users in our data set is 26 years old with a standard deviation of 10, suggesting many young users in Facebook. For each age group a, let S_a be the set of status updates written by users from that age group. We calculate the probability of emotion expression for each age group a as $P(S_i|S_a) = \frac{|S_i \cap S_a|}{|S_a|}$ with S_i (for $i = 0, 1, \ldots, 10$) defined as in the beginning of Section 4.3. Based on Figure 4.3, the probability of expression of emotions increases with age. Users post more positive emotions as they get older. We find that older users are more emotional

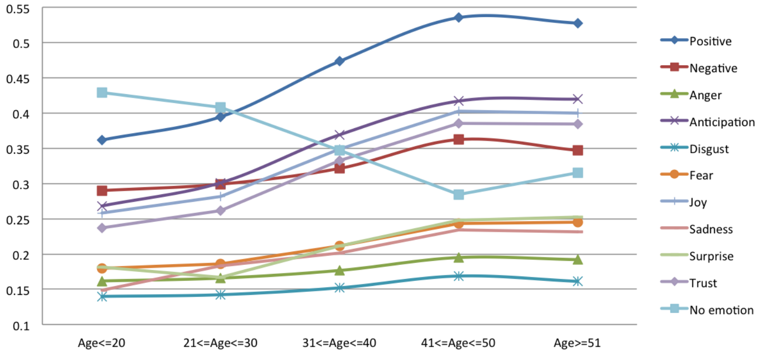


Figure 4.3: Probability of occurrences of emotions in status updates from users of different age groups. Users are more likely to post emotions as they get older.

in their posts compared to younger users. Users between 40 to 50 years old have the smallest amount of status updates without emotion expression (less than 30%), which indicates their willingness to share their feelings. On the other hand, more than 40% of young users' posts (users less than 21 years old) are without emotions. This evidence could be caused by their language use and the fact that our dictionary does not contain all possible expressions.

4.3.3 Emotion and Time

In this section, we investigate the relation between emotion expression and the time stamp of the posts. The graphs in this section depict the conditional probabilities of emotion expression w.r.t. time using $P(S_i|S_t) = \frac{|S_i \cap S_t|}{|S_t|}$, where S_t is the set of status updates posted in a specific time interval. In Figures 4.4, there are 7 such time intervals, each one corresponding to a day of the week and the time intervals for months correspond to the months of the year.

- *Emotion and day of the week:* Facebook status updates are most likely to convey emotions on Thursday. From Friday onwards, the probability of emotion expression decreases. On Saturdays, users are least likely to express any emotions in their posts. Interestingly, the frequency of status updates conveying anger and surprise remains constant from Monday to Thursday. However, on Friday, users express more surprise and become less angry in their posts. In addition, users are more negative during the workdays and less likely express to joy. However, on Saturday and Sunday, users become less negative and more joyful. Figure 4.4 presents that people are more emotional

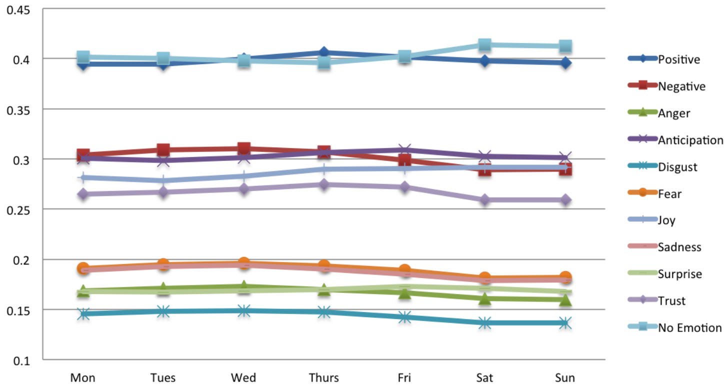


Figure 4.4: Probability of occurrences of emotions in status updates depending on the day of the week. Status updates are more likely to contain emotions during workdays than during the weekend.

during workdays than weekends. During the weekend (on Saturday and Sunday), users are less emotional and their posts are more likely without emotion expression. Among other things, the number of posts about trust decreases during the weekend, and a similar observation holds for posts related to fear.

- *Emotion and month of the year:* Figure 4.5 presents the probability of emotion expression during different months of the year. Facebook users are more emotional in December; in particular, users are less negative, more joyful, surprised, anticipating and positive compared to other months of the year. This is reflected in posts such as *"Happy holiday", "Happy NYE", "Happy Christmas"* which are very prominent in December and which are tagged as emotion conveying posts by the emotion detection method described in Section 4.2. Although there are no significant changes in emotions during the rest of the year, during the summer months (June, July and August), the amount of positive, fear and trust expressions decreases, and users' posts are least likely to contain any emotion.

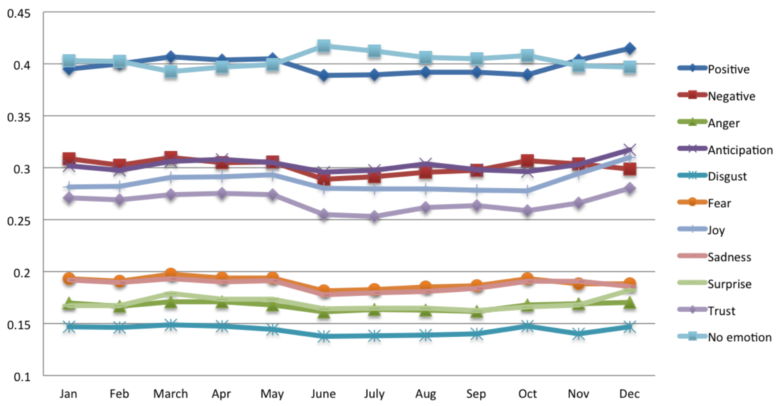


Figure 4.5: Probability of occurrences of emotions in status updates depending on the day of the week. Status updates are more likely to contain emotions during workdays than during the weekend.

4.3.4 Emotion and Personality

Similarly as in the previous sections, for each of the personality traits, we consider the set of status updates written by users who meet the threshold for that personality trait according to Table 3.2. Using S_p to denote the set of status updates linked in this way to personality trait p, we compute $P(S_p|S_i) = \frac{|S_i \cap S_p|}{|S_i|}$ with S_i (for $i = 0, 1, \ldots, 10$) defined as in the beginning of Section 4.3. The results are visualized in Figure 4.6. Similarly, results of $P(\neg S_p|S_i) = \frac{|S_i \cap \neg S_p|}{|S_i|}$ are visualized in Figure 4.7. Neurotic users' posts are

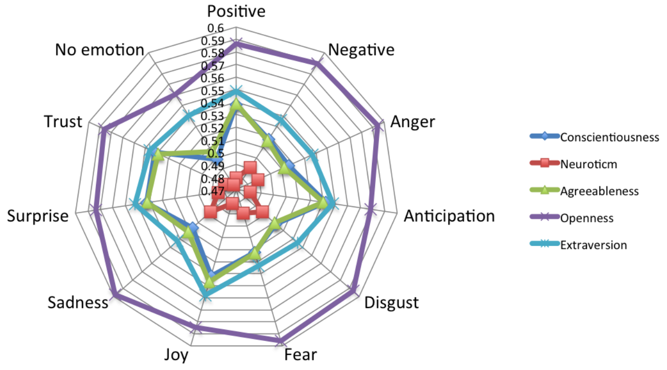


Figure 4.6: Probability of occurrences of emotions in status updates from users of different age groups. Users are more likely to post emotions as they get older.

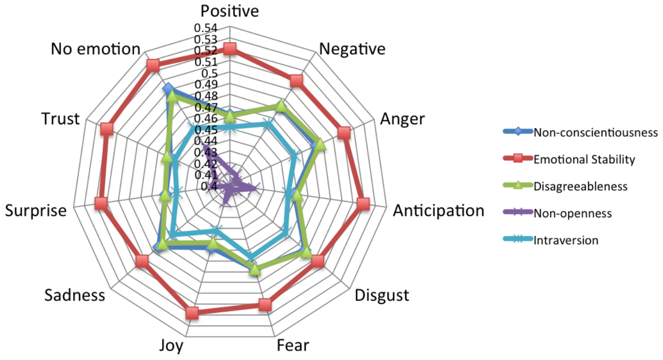


Figure 4.7: Probability of occurrences of emotions in status updates from users of different age groups. Users are more likely to post emotions as they get older.

less likely to be emotional, while open users' posts convey emotions more frequently than other personalities. After open users, extrovert users express the most emotions in their posts. Interestingly, agreeable users express emotions very similar to conscientious users on Facebook.

Posts containing anticipation are mostly expressed by agreeable, conscientious and extrovert users. Neurotic users use less joy expressions than other personalities and their posts are most likely about disgust, sadness and negative feelings. Sadness appears more than other emotions for neurotic and open users, while joy emotions are expressed most by extrovert, conscientious and agreeable users. Open users also post frequently about their fear and anger.

4.4 Conclusions and Future Work

In this study, we explored the relation between the emotions of 5,865 Facebook users with their age, gender and personality by using their status updates (almost 1 million posts). We used the NRC hash-tag emotion lexicon to detect emotions from the posts. We also extracted temporal features from the posts' time stamps. Almost 60% of status updates contain at least one type of emotion expression. Positive emotion is expressed with the highest frequency in status updates and disgust is least likely to appear in the status updates of users.

The results confirm a relation between users' characteristics and their emotions. Similar to offline expression, female Facebook users express more emotions in their status updates than male users. Similarly, older users express more emotions in their status updates than younger users. Neurotic users are not very emotional in their status updates, while open users are mostly likely to express their feelings about different subjects. By analyzing the time stamp of the status updates, we examined relations between Facebook posts' time and users' feelings. Interestingly, emotions are more likely to be expressed during the workdays compared to the weekend. The frequency of emotional status updates is lowest during the summer and highest in December.

We found significant correlations between our selected features and users' emotions. In future research, we envision developing a model that will predict the most probable upcoming emotion for each user, among other things based on time, demographics and personality. We believe that being able to predict users' emotions and target the end users accordingly would be useful for personalized services.

Aside from the work we have presented in this work, there is clear potential for more fine grained emotion detectors. Emotion detection in this study has been performed using a lexicon based approach. However, due to the complexity of the status updates, the limited size of the lexicon, and a huge amount of noise in the unnormalized status updates, it is very likely that we have missed many emotion expressions in the status messages. Exploring better techniques to extract emotions not only based on the words, but also based on other features is potentially an open path to explore.

Chapter 5

REGRESSION APPROACH TO BIG 5 PERSONALITY RECOGNITION

In this chapter we present the details of the experiments and the results of personality prediction using two different datasets described in Chapter 3.

5.1 Experimental Setup

We use the R software environment for running different prediction models (Support Vector Regressors and Linear Regression) and computing Spearman and Pearson correlations between the features and the five personality scores. R is a language and environment for statistical computing and graphics. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques and is highly extensible [63]. We also use the Mulan java library for running models for multi-target predictions. Mulan is an open-source Java library for learning from multi-label datasets [71]. The goal of our personality recognition task is to predict a personality profile, in the form of 5 scores, one for each of the 5 traits (*Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness* and *Emotional Stability*), for a given user.

Formally, multivariate regression addresses the following problem:

Let F be the vector (feature space or input space) including m features, $f_1, f_2, ..., f_m$ (described in Chapter 3), and T be the target vector (output space) including 5 target variables $t_1(Openness)$, $t_2(Conscientiousness)$, $t_3(Extraversion)$, $t_4(Agreeableness)$ and $t_5(Neuroticm$ or $Emotional\ Stability)$. The goal of a multivariate regression algorithm is to learn a model $M : F \rightarrow T$ that minimizes the prediction error($RMSE$, described in Chapter 2.2) over a test set. The goal of a univariate regression algorithm is to learn 5 models

$M_1 : F \rightarrow t_1$ $(Openness)$

$M_2 : F \rightarrow t_2$ $(Conscientiousness)$

$M_3 : F \rightarrow t_3$ $(Extraversion)$
$M_4 : F \rightarrow t_4$ $(Agreeableness)$
$M_5 : F \rightarrow t_5$ $(Neuroticm)$
that minimizes the prediction error ($RMSE$, described in Chapter 2.2) over a test set.

5.2 Experiments on YouTube vloggers dataset

5.2.1 Introduction

In this section we focus on multimodal personality impression recognition of video bloggers (vloggers). Analysis of video content appears to be one of the least studied problems in the domain of computational personality recognition [9]. The work done here is different from similar works in this domain [28, 22, 6], in the sense that the ground truth data does not come from the vloggers themselves, but from other users watching the videos. In other words, the task that we address is not recognition of the true personality traits of vloggers, but *predicting how the personality of vloggers is perceived by their viewers.* To this extent we use both non-verbal cues, i.e. audio-video features, as well as textual analysis of the transcripts of the videos (refer Chapter 3.3).

Given a video, the aim is to obtain 5 scores based on the 10-item measure of the Big Five[1] (or Five-Factor Model) dimensions. We treat this problem both as five univariate (in which we model each personality trait independently) as well as a multivariate regression task (in which we make a combined prediction for all 5 personality trait scores, instead of training a regressor for each trait separately) as decsribed in Section 5.1. Some initial research has been done on the use of multivariate regression for personality prediction on Facebook [5, 36] and Sina Microblog [7]. In the current section we investigate whether the promising trend of good results can be extended to perceived personality prediction of vloggers. In particular, we measure the performance of 5 multivariate regression techniques [74], namely *Multi-Target Stacking*(MTS), *Multi-Target Stacking Corrected*($MTSC$), *Ensemble of Regressor Chains*(ERC), *Ensemble of Regressor Chains Corrected*($ERCC$) and *Multi Object Random Forest*($MORF$)(described in Chapter 2.1.2), on a YouTube personality dataset [9](described

[1] http://www.sjdm.org/dmidi/Ten_Item_Personality_Inventory.html

in Chapter 3.3). We contrast these 5 multivariate regression techniques with univariate approaches such as correlation based feature selection algorithm, as well as a single target approach using decision trees and SVM and a mean baseline algorithm.

5.2.2 Experiments

Using the univariate and multivariate regression formulation defined in Section 5.1, we present the results of different models as enumerated in this section. We compared Support Vector Regressor (SVM) [39] and Linear Regression (LR) as our learning models for the univariate approach, but since SVM models outperformed LR, we present the results of only SVM here. We used SVM with rbfdot kernel and regularization constant (C) set to the default value of 1 in the R software. We evaluated models with different choice of kernels like linear, polynomial, but since the rbfdot kernel consistently outperformed the other kernels, we present the results with only rbfdot kernel. Mulan is used to run models on all the multivariate regression algorithms listed above. The WCPR 2014 workshop organisers [15, 9] provided us with a training (348 vloggers) and test (56 vloggers) split which is used in all experiments. The baseline is the model that returns the mean score for each personality trait (refer Table 3.4 for the mean scores). We use root mean squared error (RMSE) and co-efficient of determination (R^2) (refer Chapter 2.2) as our evaluation criteria. To measure significant differences in prediction errors between the learned models and the baseline, we conducted two-tailed paired t-tests for the RMSE, and two-tailed single t-tests for R^2.

1. **Correlation based Feature Selection using Forward Search**

   In this model, we build 5 univariate SVM regression models for each of the 5 personality traits wherein each personality trait model is trained using a differet feature set combination which is computed using correlation based forward search technique. The algorithm for forward search technique is explained in detail in Chapter 2.3.1. This technique is used to find the best feature set combination for each trait, which minimises the $RMSE$ (objective function), in the given training dataset using 10 fold cross validation. We then evaluate the obtained feature set combination against the test dataset. The feature sets that are coded in the result tables (Table 5.2) are described in Table 5.1. The letter codes are combined for experiments where different feature sets were used together (e.g. AV+cL is a system trained on all Audio/Video and correlated LIWC features).

Table 5.1: Forward search: Feature set category and the corresponding letter codes.

Feature Set	Letter Code
Audio Video and Gender	AV
(Correlated) Audio Video and Gender	cAV
LIWC	L
(Correlated) LIWC	cL
NRC Emotion	N
(Correlated) NRC Emotion	cN
Senti Strength	SS
(Correlated) Senti Strength	cSS
MRC	M
(Correlated) MRC	cM
Splice	S
(Correlated) Splice	cS

By employing a combination of wrapper and filter based feature subset selection, we begin forward search by finding Spearman significantly correlated features ($p < .05$), for each feature set category (as in Table 5.1) and each personality trait on the training dataset. Using 10 fold cross validation, $RMSE$ values are computed for all and Spearman significantly correlated features for each feature set and for each trait independently (refer Table 5.2). We repeat this computation on the 6 different feature sets (see the appendix for the full list of significantly correlated features for each personality type). We build the Table 5.2 using the below rules. Feature sets without any significantly correlated features with the personality trait are indicated as NA. Ranks are assigned to each feature set as follows:

(a) Within each feature set category, either all (e.g. AV) or correlated features (e.g. cAV) are selected based on the $RMSE$ value. The other feature sets will not be considered in the further steps and are marked as $-$.

(b) $RMSE$ values above the baseline are not considered and marked as $-$.

(c) The lowest $RMSE$ value gets the highest rank (1 being the highest).

(d) Feature sets with the same $RMSE$ value, get the same ranks.

The best performing feature set for each trait is indicated in bold in Table 5.2. Beginning with the highest rank, feature sets are combined, if and only if the $RMSE$ value of this combination is lower or remains unchanged. If not, the combining feature set is dropped from the next iteration. We continue this process until all the feature sets are processed. Table 5.3 presents the best feature set combination for each of the 5 personality traits obtained using this technique. Using this feature set combination, we train five univariate regression models using SVM on the entire training dataset of #348 vloggers and predict the scores on the #56 test vloggers dataset. Results are presented in Table 5.4.

2. **Stacked Forward Search**

   In this univariate model , we augment the feature space of the best feature set combination obtained in Table 5.3 using the scores of the other 4 personality traits. In particular, we use actual personality scores for training our model on the #348 vloggers dataset and predicted scores, obtained using the Forward Search described above, on the #56 test vloggers dataset to obtain the final set of prediction scores. We run our models using SVM regression. Our approach is similar to the Multi Target Stacking Corrected ($MTSC$) algorithm explained in Chapter 2.1.2 but we use a unique feature set combination for each trait based on its performance on the training dataset.

3. **All and Correlated features**

   In this univariate model, we use all audio/video, gender and linguistic features to train and test our models. We alo use the Spearman significantly correlated features

Table 5.2: RMSE Comparison on #348 training on a YouTube vloggers dataset using all and correlated features under each feature set category. All results are based on 10 fold cross validation using SVM (rbf kernel, C=1).

	Extr	Rank	Agr	Rank	Cons	Rank	Emo-Stab	Rank	Open	Rank
	(E)	(E)	(A)	(A)	(C)	(C)	(ES)	(ES)	(O)	(O)
AvgBaseline	.97		.86		.78		.79		.70	
AV	**.83**	**1**	.90	–	.76	4	.81	–	**.68**	**1**
cAV	.85	–	.87	–	.77	–	.81	–	.71	–
L	.89	2	.78	–	**.71**	**1**	.74	–	**.68**	**1**
cL	.91	–	**.75**	**1**	**.71**	**1**	**.72**	**1**	**.68**	**1**
N	.98	–	.82	2	.76	4	.75	2	.70	3
cN	NA	–	.87	–	.79	–	.79	–	.72	–
SS	1	–	.84	–	.80	–	.75	2	.71	–
cSS	NA	–	.83	3	.79	–	.76	–	.69	2
M	.96	–	.90	–	.75	–	.75	2	.72	–
cM	.95	3	NA	–	.72	2	NA	–	NA	–
S	.96	4	.86	–	.76	–	.76	3	.71	–
cS	.97	–	.84	4	.75	3	.77	–	.7	3

Table 5.3: Best feature set combination obtained using Forward Search on the #348 training on a YouTube vloggers dataset for the 5 traits: *Extraversion (Extr)*, *Agreeableness (Agr)*, *Conscientiousness (Cons)*, *Emotional Stability (EmoStab)*, *Openness (Open)*. All results are based on 10 fold cross validation using SVM (rbf kernel, C=1)

	Avg Baseline	Best Feature Set Combination	RMSE (Best Feature Set)
Extr	.97	AV+cM	.82
Agr	.86	cL+N+cSS+cS	.75
Cons	.78	cL+cM+cS+AV	.71
EmoStab	.79	cL+M+N	.71
Open	.70	AV+L+N+cS	.67

($p < .05$) for each personality trait, computed on the training dataset, to train and test our models. We run our models using SVM regression.

4. **Linguistic and Correlated linguistic features**
   We use only the linguistic features (LIWC, NRC, MRC, Senti Strength and Splice) computed from the vloggers' transcripts and Spearman significantly correlated linguistic features ($p < .05$) for each personality trait, computed on the training dataset, to train and test our univariate models using SVM regression.

5. **Multivariate Regression using Mulan**
   We use the Mulan implementation of the 5 multivariate regression algorithms (MTS, MTSC, ERC, ERCC and MORF) and Single Target (ST) regression to obtain our results.[2] Note that ST does not leverage the prediction result for one personality trait to make a prediction for another, while all other algorithms do in one way or another. All algorithms except MORF use Weka decision trees as a base learner. For detailed explanation about these algorithms, please refer to Chapter 2.1.2. For further information we refer to [74].

5.2.3 Performance Comparison on YouTube vloggers dataset

It can be seen from the results in Table 5.2 that audio video and LIWC features produce the best models in the training dataset for all 5 traits when used independently. But the best feature set combination is unique for each trait since the feature interactions that occur when learning the model for each trait is different. It can be seen from the results in Table 5.4 that all 6 algorithms (ST, MTS, MTSC, ERC, ERCC and MORF) as well as SVM models using different feature set combinations outperform (i.e., have a lower prediction error than) the baseline model for all 5 personality types. In addition, positive values for R^2 are also observed for all (except correlated linguistic feature combination for *Openness* trait and stacked forward search approach for *Emoional Stability*) the models which further indicates better performance than the average baseline model ($0\% \leq R^2 \leq 37\%$).

[2] http://mulan.sourceforge.net/

Table 5.4: Root mean square error ($RMSE$) and Coefficient of determination (R^2) results for personality trait prediction using *univariate* and *multivariate* regression algorithms on a YouTube vloggers dataset. In each column, the lowest error and highest determination are typeset in bold. Significant differences ($p < .05$) are marked using $*$.

	Extr		Agr		Con		EmoStab		Open	
Feature Set	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2
Average Baseline	1.02		.91		.71		.75		.83	
Multivariate Regression in Mulan										
MTS	.95	13	.76	30	.66*	14*	.71	10	.77	14
MTSC	.91	20	.73	35	**.64**	**18**	**.70**	**13**	.79	9
ERC	.93	17	.74	34	.65*	16*	.72	8	.79*	9*
ERCC	.91	20	**.72**	**37**	.65	16	**.70**	**13**	.80	7
MORF	.98	7	.84	15	**.64**	**18**	.75	0	.83	0
Univariate Regression using decision trees in Mulan										
ST	.91	20	**.72**	**37**	.65	16	**.70**	**13**	.80	7
Univariate Regression using SVM with rbf kernel										
Forward Search	**.88***	**26***	.77	28	.68	8	.75	0	.76	16
Forward Search (with Stacking)	**.88***	**26***	.78	27	.68	8	.79	-11	**.74***	**21***
All Features	.91	20	.77	28	.67	11	.72	8	.8	7
Correlated features	**.87***	**27***	.77	28	.65	16	.74	3	.82	2
Linguistic Features	.95	13	.76	30	.68	8	.73	5	.83	0
Correlated Linguistic	.92	19	.77	28	.67	11	.74	3	.84	-2

For *Extraversion*, correlated features works the best followed by forward search (stacked and unstacked) using SVM among all the other systems ($R^2 = 27\%$, $R^2 = 26\%$ and $R^2 = 26\%$ respectively). Interestingly, correlation based feature selection has a positive effect only for this trait other than *Conscientiousness.*

For *Agreeableness*, ST and ERCC are the best performers ($R^2 = 37\%$). It is interesting to note that correlated features did not show any improvement in RMSE value compared to all features. Further study into possible reasons for this behaviour in some traits is an interesting area of research in the future. Using decision trees as the base learner in ST and ERCC performs better than the SVM model for this trait.

In the case of *Conscientiousness*, MTSC and MORF emerges as the best models with lowest prediction errors ($R^2 = 16\%$). Interestingly, MORF outperformes for this personality type, but performs poorly for the others. In particular, it fails to fit the data correctly for Emotional Stability and Openness ($R^2 = 0\%$). Forward search technique did not show any improvement for this trait, whereas using correlated and correlated linguistic features shows improvement compared to all and linguistic feature sets.

Emotional Stability has three main winners – ST, MTSC and ERCC ($R^2 = 13\%$). It is interesting to note that correlation based feature selection has a negative effect for this trait as well as *Openness*. One possible reason could be that the correlation measure is not sufficient to identify relevant features in isolation due to high feature interaction (for more information on correlation based feature selection we refer to [32]).

Finally, in case of the *Openness* personality type, forward search with stacking has the lowest prediction error ($R^2 = 21\%$) when compared to the baseline as well as all other models. The forward search based models results for *Extraversion* and *Openness* traits indicate that each personality trait works better with its unique feature set rather than one feature space for all traits.

An interesting observation is that, in past studies, models for Extraversion personality trait often performed best, while models for Agreeableness performed worst (e.g.[5], [9]). However, our results show that models for Agreeableness as well as Extraversion were the top performers among the 5 personality types. The prime difference in the settings of the previous studies (mentioned above) and ours are: the set of features used (in both cases), the dataset (in [5]), among others. Therefore, it is hard to put a finger on the exact cause for this change in the performance. This could be an interesting problem to investigate in future. Finally, our overall prediction results are important because previously published methods [9] for the same dataset show an improvement over the baseline for the majority of personality traits, but not for "all". Furthermore, based on our results we can say that multivariate regression, in combination with feature selection, does show potential for solving personality prediction problems. However, it is possible for single target methods to give equally good results.

5.3 Experiments on myPersonality Dataset

5.3.1 Introduction

In this section we focus on Big 5 personality score prediction of Facebook users. For this purpose we use user demographic features (age and gender) along with measures that are computed from the user's Facebook profile (e.g #groups, #network size, #likes and #education). We combine all the status posts made by a given user as one document for each user. Thus, we have a total of 38,106 documents aggregated from approximately 7 million status updates of 38,106 users in our dataset. It has been shown in psychological studies [44] that there exist links beween linguistic features (extracted from text and conversation) and users' personality traits, which are demonstrated using correlations on acoustic parameters, lexical categories, ngrams etc [60]. Thus it is increasingly popular to use language in social media for predicting personality. These findings motivate the choice of linguistic features that we use in our experiments. We follow a traditional text analysis technique which is very widely used in psychology studies namely the Linguistic Inquiry and Word Count (LIWC). MRC features used in previous studies [27] showed significant correlations between features like concreteness and *Extraversion*. *Conscientiousness* is also shown to be associated with words expressing insight, longer words (Nphon, Nlet, Nsyl and Sixltr) as well as words that are acquired late by children (AOA) in the MRC database. Additionally, we also include 2 sentiment features (Positive and Negative sentiment) using the Senti-Strength tool to our feature space, since many studies have successfully exploited emotion and sentiment features in personality prediction tasks [15]. Finally, we use 71 features obtained using the Splice API which includes Part of Speech Cues, the Immediacy cues, and the Tense cues.

Given a set of Facebook profile features and linguistic features extracted from the combined status text for each user, the aim is to obtain 5 personality scores based on the 20-336 item IPIP proxy for Costa and McCrae's NEO-PI-R domains (Five Factor Model) [19]. In the current section we investigate whether the promising trend of good results of multivariate regression on Youtube vloggers can be extended to personality prediction of Facebook users. In particular, we measure the performance of 5 multivariate regression techniques [74], namely *Multi-Target Stacking*(MTS), *Multi-Target Stacking Corrected*($MTSC$), *Ensemble of Regressor Chains*(ERC), *Ensemble of Regressor Chains Corrected*($ERCC$) and *Multi Object Random Forest*($MORF$) (described in Chapter 2.1.2), on a myPersonality Facebook dataset (described in Chapter 3.1.2). We contrast these 5 multivariate regression techniques with univariate approaches such as correlation based feature selection algorithm, as well as a single target approach using decision trees and linear regression and a mean baseline algorithm.

5.3.2 Experiments

Using the univariate and multivariate regression formulation defined in Section 5.1, we present the results of different models as enumerated in this section. The baseline is the model that returns mean score for each personality trait(refer Table 3.3 for the mean scores). We use Linear Regression(LR) as our learning algorithm for the univariate approach. Mulan is used to run models on all multivariate regression algorithms listed above. We use root mean squared error ($RMSE$) and co-efficient of determination (R^2) as our evaluation criteria. All $RMSE$ values are averaged over 10 fold cross validation on the entire dataset of #38,106 users. We also measure Pearson correlations between the predicted scores and the actual values. A positive value indicates a similar and identical relation between the actual and predicted scores.

1. **All and Correlated Features**

   We use all the features (demographic, FB Activity and linguistic features) to train and test our univariate linear regression model. For each fold, we use Pearson Correlation to compute significantly correlated features ($p < .05$) on the training data for each trait. We use the obtained correlated features to train our model on each fold. Using this learning model, we predict the scores on the test dataset on that fold. Thus in each fold, we compute the correlation of train data without any information outside of the training data, making the test data an out-of-sample evaluation of the obtained feature space.

2. **Stacking Approach**

   In this univariate linear regression approach we augment the feature space for a given personality type by using the scores of the other 4 traits. We use linear regression as our learning algorithm. We begin by randomly splitting the dataset into 10 folds and in each fold, we train models for the other 4 personality types using the entire feature space. Predicted scores are obtained for the other 4 personality types on the test dataset for that fold. We augment the feature space for that fold using the other 4 actual personality scores on the training dataset and predicted scores on the test dataset. For each fold, models are built on the augmented training dataset and tested on the augmented test dataset. We repeat this method for all 5 traits and on all 10 folds to obtain the final average $RMSE$ values.

3. **Linguistic and Correlated linguistic features**

   We use only the linguistic (LIWC, MRC, Senti Strength and Splice features) and Pearson significantly correlated linguistic features ($p < .05$) to train and test our univariate linear regression models. We use the same approach as described above to compute correlation for each fold.

4. **Multivariate Regression using Mulan**

   We use the Mulan implementation of the 5 multivariate regression algorithms (MTS, MTSC, ERC, ERCC and MORF) and Single Target (ST) regression (refer Chapter 2.1.2) to obtain our results.[3] Since Mulan library is not designed for handling huge datasets like ours, we run our models on Mulan using a random sample of #3400 users from our #38,106 dataset.

[3] http://mulan.sourceforge.net/

Table 5.5: Root mean square error (RMSE) results for personality trait prediction on #38,106 users of myPersonality dataset using all features under each feature set. All values are averaged over 10 fold cross validation using linear regression. In each column, the lowest error are typeset in bold.

Feature Sets	Extr	Agr	Cons	Neu	Open
AvgBaseline	.81	.70	.73	.80	.67
Demog/FB Activity	**.79**	.692	**.715**	**.783**	.663
LIWC	.80	**.687**	.719	.796	**.652**
Senti-Strength	.802	.69	.727	.80	.66
MRC	.804	.692	.725	.80	.66
Splice	.799	.691	.722	.798	.66

5.3.3 Performance Comparison on myPersonality Dataset

Regression results for different feature sets are compared in Table 5.5. Since using correlated and all the features under each feature set category did not show much difference in the RMSE value, we present the results using all features under each feature set. See the appendix for the full list of Pearson significantly correlated features ($p < .05$) for each personality type.

For *Extraversion*, all the multivariate regression models except MORF produce an almost similar $RMSE$ value of .78 ($R^2 = 7.89\%$), with very small differences in the individual values. The stacking approach did not show any improvement over the unstacked approach. Demographics and FB Activity features work the best, producing the lowest RMSE of .79 when modelled independently. But when linguistic features are included to this feature space, we see an improvement in the error rate to .78. The Pearson Correlation co-efficient between the predicted scores and the actual values for this trait is .26 ($p < .001$). For all the traits, it can be observed that compared to other feature sets, LIWC and Demographics/FB Activity features on their own perform the best.

For *Agreeableness*, the stacking approach with all features is the winning model with a RMSE value of .68 ($R^2 = 5.35\%$). It is interesting to note that all the multivariate regression models produce a very similar $RMSE$ of .69. For all the 5 personality traits, we can

Table 5.6: Root mean square error ($RMSE$) and Coefficient of determination (R^2) results for the personality trait prediction using *univariate* linear regression and *multivariate* regression algorithms on #38,106 and #3400 users respectively on a myPersonality dataset. In each column, the lowest error and highest determination are typeset in bold.

	Extr		Agr		Con		Neu		Open	
Feature Set	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2	*RMSE*	R^2
Avg Baseline	.81		.70		.73		.80		.67	
Multivariate Regression in Mulan										
MTS	.782	6.79	.698	.6	.717	3.53	.772	6.89	.65	5.88
MTSC	.777	7.98	.69	2.83	.714	4.34	**.763**	**9.04**	.649	6.17
ERC	**.776**	**8.22**	.69	2.83	.713	4.6	.766	8.32	.649	6.17
ERCC	**.776**	**8.22**	.69	2.83	.713	4.6	**.763**	**9.04**	.649	6.17
MORF	.787	5.6	.693	2	.72	2.72	.774	6.39	.653	5.01
Univariate Regression using decision trees in Mulan										
ST	.777	8	.691	2.55	.713	4.6	.765	8.56	.649	6.17
Univariate Regression using least squares linear regression										
All Features	.781	7.03	.685	4.24	.708	5.94	.779	5.18	**.638**	**9.32**
All Features (with Stacking)	.780	7.27	**.681**	**5.35**	**.703**	**7.26**	.776	2.53	.643	7.9
Correlated (features)	.781	7.03	.685	4.24	.707	6.2	.78	4.93	.646	7.04
Linguistic (features)	.795	3.67	.685	4.24	.716	3.8	.794	1.49	.649	6.17
Correlated linguistic	.796	3.43	.686	3.96	.716	3.8	.796	1	.648	6.46

observe that Single Target (ST) outperforms the Multi-Target Stacking (MTS) algorithm. In terms of correlation between the predicted scores and the actual values, the correlation of .17 ($p < .001$) is observed.

For *Conscientiousness*, using all the features with the stacking approach is the best performing model with a RMSE value of .703 ($R^2 = 7.26\%$). Similar to extraversion, this trait produced significant Pearson correlation of .25 ($p < .001$) with the actual scores.

In case of *Neuroticm*, MTSC and ERCC produce the best models with RME of .763 ($R^2 = 9.04\%$). It can be seen from the YouTube vloggers results, that MTS and ERCC were the best performing model for Emotional Stability. Furthermore, feature selection did not work well for this trait. Similar observations can be made for this dataset too. Decision trees used in MTS and ERCC turns out to be the winning algorithm for this trait.
Finally for *Openness*, the linear regression model using all the features is the best model with the lowest RMSE value of .638 ($R^2 = 9.32\%$). In terms of correlation between the predicted scores and the actual values, the linear regression model produce significant Pearson Correlation co-efficient of .24 ($p < .001$).

It is important to note that given a strong baseline RMSE of .742, averaged over all 5 traits, we are still able to outperform it by achieving an average RMSE of .712. Recall that for the YouTube vloggers dataset, where the average baseline RMSE is .844 and using our models we obtain an average RMSE of .74.

5.4 Conclusions and Future Work

In this work, we explored different approaches to computational personality recognition of social media users. We predict personality on a continuous scale common in psychology studies. While we predict the perceived personality scores from spoken text (transcripts from video) in YouTube vloggers dataset, we predict the self reported pesonality scores from writtern text (status updates) in myPersonality dataset. Additionally, we use other features specific to the 2 datasets like the audio video features and Facebook profile features for the vloggers and Facebook users respectively. Table 5.7 summarises the best personality recognition models and feature sets we used. In the case of Youtube vloggers dataset, we explored SVM models for univariate regression in which we use correlation based feature selection using forward search. In the case of Facebook users' dataset, we explored linear regression models for the univariate approach in which we compared correlated features with the entire feature space. In addition, using the fact that the personality traits are

Table 5.7: Comparison of the best model for the 5 traits: *Extraversion (Extr)*, *Agreeableness (Agr)*, *Conscientiousness (Cons)*, *Emotional Stability (EmoStab)*, *Openness (Open)* on the myPersonality and the YouTube vloggers dataset. Each row consists of algorithm, feature set used and the obtained RMSE improvements measure from the mean baseline. See Section 5.2 and Section 5.3 for details.

Task	Regression		
	Algorithm	Feature Set	Model Performance
Avg Baseline	-	-	0%
Self Report Personality Models of Facebook users			
Extr	ERC and ERCC	All	4.2%
Agr	LR	Stacking Aproach	2.7%
Cons	LR	Stacking Aproach	3.7%
Neu	MTSC/ERCC	All	4.6%
Open	LR	All	4.8%
Perceived Personality Models of YouTube vloggers			
Extr	SVM	Correlated Features	13.7%
Agr	ERCC and ST	All	21%
Cons	MTSC and MORF	All	9.9%
Emo Stab	MTSC and ERCC	All	6.7%
Open	SVM	Stacked Forward Search	10.8%

highly correlated, we also augment the feature space using the predicted scores of the other 4 personality traits. We also experiment univariate models using different feature spaces, like linguistic or correlated linguistic features. We contrast this approach with multivariate regression by employing 5 different algorithms. Note that all the multivariate models (except $MORF$) are build using decision trees as the base learner whereas univariate approaches use SVM and LR. We observe that models for perceived personality traits outperform models for self reported personality, consistent with results observed in [44]. Infact our results reveal the state of the art average $RMSE$ of .76 in computational personality recognition from multimodal features for perceived personality impression scores of YouTube vloggers dataset [15]. While *Extraversion* followed by *Openness* are the easiest to predict by the observers in voggers, *Openness* followed by *Neuroticm* are the best performing traits in the self reported personality models of Facebook users.

In the case of myPersonality dataset, we could not apply forward search, since all the feature subsets had produced identical $RMSE$ values and hence could not apply stepwise feature addition. In the future, we would like to investigate other measures for computing the correlation between features and the personality scores like *information gain*, in addition to linear correlation measures. Since the myPersonality dataset is huge(#38,106 users) to be handled by the SVM algorithm [39], running models on the SVM was very time consuming. Hence we report our resuls using LR only. Another issue is that, several linguistic features had shown significant correlations with the personality traits in the myPersonality dataset, but did not seem to improve the performance of the model compared to using all liguistic features. Research into possible reasons for such behavior(e.g. varying levels of feature interaction that could exist within the feature space) need to be studied. Using feature selection techniques, like correlation based forward search, as well as using a different base learner like SVM/LR in multivariate algorithms will be an interetsing area to research in the future.

Chapter 6
CONCLUSION

6.1 Summary

In this study, we show that users' personality can be recognised from several different features from their text or video transcripts, in addition to other measures. We explore the use of different univariate regression techniques like Linear Regression and Support Vector Regression as well as multivariate regression algorithms to predict the self reported personality scores of Facebook users and perceived personality scores of YouTube vloggers. Instead of training 5 learners separately to predict the 5 personality scores, multivariate regression techniques make a combined prediction of the 5 personality scores. Given the correlation among different personality traits, this sounds promising. We observe that no common learning algorithm and feature space works well for all the five traits. Each personality trait has its unique learning algorithm and feature set that has worked well.

For the YouTube vloggers dataset, although the multivariate regression techniques that we evaluated performed well on a YouTube personality dataset of 404 vlogs, they did not clearly outperform a single target approach in which a model was trained for each personality trait separately. Correlation based forward search outperformed the other methods for some traits, hence feature selection is very important since these models use only a subset of the full feature space. All the models developed outperformed our average prediction baseline though for all 5 personality dimensions. Past methods proposed for the same dataset were able to outperform the baseline for 3 personality traits simultaneously only [9]. Infact our results reveal the state of the art average $RMSE$ of .76 in computational personality recognition from multimodal features for this dataset [15]. For the myPersonality dataset, even though we improve over the baseline $RMSE$ using all our different models, we did not achieve significant improvement over the baseline for any of the 5 traits. The average baseline score seems to be a strong baseline for this dataset since the numeric scores are very close to the mean score. Hence we are not able to significantly outperform this mean baseline. Exploring better feature selection techniques and learning algorithms is a direction of future research.

Additionally, we explored the relation between the emotions of 5,865 Facebook users with their age, gender and personality by using their status updates (almost 1 million posts). We used the NRC hash-tag emotion lexicon to detect emotions from the posts. We also extracted temporal features from the posts' time stamps. Almost 60% of status updates contain at least one type of emotion expression. Positive emotion is expressed with the highest frequency in status updates and disgust is least likely to appear in the status updates of users.

The results confirm a relation between users' characteristics and their emotions. Similar to offline expression, female Facebook users express more emotions in their status updates than male users. Similarly, older users express more emotions in their status updates than younger users. Neurotic users are not very emotional in their status updates, while open users are mostly likely to express their feelings about different subjects. By analyzing the time stamp of the status updates, we examined relations between Facebook posts' time and users' feelings. Interestingly, emotions are more likely to be expressed during the workdays compared to the weekend. The frequency of emotional status updates is lowest during the summer and highest in December.

6.2 Future Work

We believe that being able to predict users' emotions and personality in order to target the end users accordingly would be useful for personalized services. Results of a study that was conducted on Amazon Mechanical Turk on 234 participants by the authors of the paper [33] concluded that advertisements are more effective when the motivational concerns of the advertisement text are congruent to the respondent's personality characterictics. We want to verify the above hypothesis in the context of social media by designing advertisements with the text tailored to an individuals personality trait. We envision developing a survey inside a Facebook Application that will enable collecting ground truth data on the users' advertisement choice as well as their profile data. By using the personality prediction models, we have the ability to infer personality scores from the user's Facebook profile. Using the knowledge of advertisement preferences from the survey, we would like to study the relationship between users' personality and their preferences for certain advertisement framing.

BIBLIOGRAPHY

[1] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.

[2] Gediminas Adomavicius and Alexander Tuzhilin. Context-aware recommender systems. In *Recommender systems handbook*, pages 217–253. Springer, 2011.

[3] Gordon W Allport and Henry S Odbert. Trait-names: A psycho-lexical study. *Psychological monographs*, 47(1):i, 1936.

[4] Ivana Anusic, Ulrich Schimmack, Rebecca T Pinkus, and Penelope Lockwood. The nature and structure of correlations among big five ratings: the halo-alpha-beta model. *Journal of Personality and Social Psychology*, 97(6):1142, 2009.

[5] Yoram Bachrach, Michal Kosinski, Thore Graepel, Pushmeet Kohli, and David Stillwell. Personality and patterns of facebook usage. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 24–32. ACM, 2012.

[6] Mitja D Back, Juliane M Stopfer, Simine Vazire, Sam Gaddis, Stefan C Schmukle, Boris Egloff, and Samuel D Gosling. Facebook profiles reflect actual personality, not self-idealization. *Psychological science*, 2010.

[7] Shuotian Bai, Bibo Hao, Ang Li, Sha Yuan, Rui Gao, and Tingshao Zhu. Predicting Big Five personality traits of microblog users. In *Proc. of IEEE/WIC/ACM WI-IAT*, volume 1, pages 501–508, 2013.

[8] M. R. Barrick and M. K. Mount. The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, (44):1–26, 1991.

[9] J Biel and Daniel Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *Multimedia, IEEE Transactions on*, 15(1):41–55, 2013.

[10] Joan-Isaac Biel, Vagia Tsiminaki, John Dines, and Daniel Gatica-Perez. Hi youtube!: personality impressions and verbal content in social video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 119–126. ACM, 2013.

[11] Avrim L Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1):245–271, 1997.

[12] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM*, 2011.

[13] Jonathan P Bowen and Silvia Filippini-Fantoni. Personalization and the web from a museum perspective. In *Museums and the Web*, volume 4, 2004.

[14] Iván Cantador, Ignacio Fernández-Tobías, Alejandro Bellogín, Michal Kosinski, and David Stillwell. Relating personality types with user preferences in multiple entertainment domains. In *Proceedings of the 1st Workshop on Emotions and Personality in Personalized Services (EMPIRE 2013), at the 21st Conference on User Modeling, Adaptation and Personalization (UMAP 2013)*, 2013.

[15] Fabio Celli, Bruno Lepri, Joan-Isaac Biel, Daniel Gatica-Perez, Giuseppe Riccardi, and Fabio Pianesi. The workshop on computational personality recognition 2014. In *Proceedings of the ACM International Conference on Multimedia*, pages 1245–1246. ACM, 2014.

[16] Olivia Chausson. Who watches what?: assessing the impact of gender and personality on film preferences, 2010.

[17] Gokul Chittaranjan, Jan Blom, and Daniel Gatica-Perez. Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, 17(3):433–450, 2013.

[18] Paul T Costa and Robert R McCrae. Normal personality assessment in clinical practice: the neo personality inventory. *Psychological assessment*, 4(1):5, 1992.

[19] Paul T Costa and Robert R McCrae. The revised NEO personality inventory (NEO-PI-R). *The SAGE Handbook Of Personality Theory And Assessment*, 2:179–198, 2008.

[20] Facebook. Company info facebook newsroom. https://newsroom.fb.com/company-info/, 2014.

[21] Golnoosh Farnadi, Geetha Sitaraman, Mehrdad Rohani, Michal Kosinski, David Stillwell, MarieFrancine Moens, Sergio Davalos, and Martine De Cock. How are you doing? Emotions and personality in Facebook. In *Proc. of EMPIRE*, pages 45–56, 2014.

[22] Golnoosh Farnadi, Susana Zoghbi, Marie-Francine Moens, and Martine De Cock. Recognising personality traits using facebook status updates. In *Proceedings of the workshop on computational personality recognition (WCPR13) at the 7th international AAAI conference on weblogs and social media (ICWSM13)*, 2013.

[23] Golnoosh Farnadi, Susana Zoghbi, MarieFrancine Moens, and Martine De Cock. Recognising personality traits using Facebook status updates. In *Proc. of WCPR*, pages 14–18, 2013.

[24] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *The Journal of machine learning research*, 4:933–969, 2003.

[25] Jui-Hsi Fu, Jui-Hung Chang, Yueh-Min Huang, and Han-Chieh Chao. A support vector regression-based prediction of students' school performance. In *Computer, Consumer and Control (IS3C), 2012 International Symposium on*, pages 84–87. IEEE, 2012.

[26] Alastair J Gill, Scott Nowson, and Jon Oberlander. What are they blogging about? Personality, topic and motivation in blogs. In *Proc. of ICWSM*, 2009.

[27] Alastair J Gill, Jon Oberlander, and Elizabeth Austin. Rating e-mail personality at zero acquaintance. *Personality and Individual Differences*, 40(3):497–507, 2006.

[28] Jennifer Golbeck, Cristina Robles, and Karen Turner. Predicting personality with social media. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 253–262. ACM, 2011.

[29] Lewis R Goldberg. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26, 1992.

[30] Lewis R Goldberg, John A Johnson, Herbert W Eber, Robert Hogan, Michael C Ashton, C Robert Cloninger, and Harrison G Gough. The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40(1):84–96, 2006.

[31] Mark A Hall. *Correlation-based feature selection for machine learning.* PhD thesis, The University of Waikato, 1999.

[32] Mark A Hall and Lloyd A Smith. Feature subset selection: a correlation based filter approach. 1997.

[33] Jacob B Hirsh, Sonia K Kang, and Galen V Bodenhausen. Personalized persuasion tailoring persuasive appeals to recipients personality traits. *Psychological science*, 23(6):578–581, 2012.

[34] Ai Thanh Ho, Ilusca LL Menezes, and Yousra Tagmouti. E-mrs: Emotion-based movie recommender system. In *Proceedings of IADIS e-Commerce Conference. USA: University of Washington Both-ell*, pages 1–8, 2006.

[35] Rong Hu and Pearl Pu. Enhancing collaborative filtering systems with personality information. In *Proc. of ACM RecSys*, pages 197–204, 2011.

[36] Francisco Iacobelli and Aron Culotta. Too Neurotic, Not Too Friendly: Structured Personality Classification on Textual Data. In *Proc of Workshop on Computational Personality Recognition, AAAI Press, Melon Park, CA*, pages 19–22, 2013.

[37] Oliver P John and Sanjay Srivastava. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research*, 2:102–138, 1999.

[38] Andreas M Kaplan and Michael Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68, 2010.

[39] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab-an s4 package for kernel methods in r. 2004.

[40] Kenji Kira and Larry A Rendell. A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning*, pages 249–256. Morgan Kaufmann Publishers Inc., 1992.

[41] Dragi Kocev, Celine Vens, Jan Struyf, and Sašo Džeroski. Ensembles of multi-objective decision trees. In *Proc. of ECML*, pages 624–631, 2007.

[42] Michal Kosinski, Yoram Bachrach, Pushmeet Kohli, David Stillwell, and Thore Graepel. Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning*, pages 1–24, 2013.

[43] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[44] François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res.(JAIR)*, 30:457–500, 2007.

[45] Georgios Paltoglou Di Cai Statistical Cybermetrics Research Group School of Computing Mike Thelwall, Kevan Buckley and Wulfruna Street Wolverhampton WV1 1SB UK Information Technology, University of Wolverhampton. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558.

[46] Gilad Mishne. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*, volume 19, 2005.

[47] Tom M Mitchell. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45, 1997.

[48] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Automatic personalization based on web usage mining. *Communications of the ACM*, 43(8):142–151, 2000.

[49] Kevin Moffit and Justin Scott Giboney. Splice.

[50] Saif Mohammad, Xiaodan Zhu, and Joel Martin. Semantic role labeling of emotions in tweets. In *Proc. of WASSA*, pages 32–41, 2014.

[51] Saif M Mohammad and Svetlana Kiritchenko. Using nuances of emotion to identify personality. *arXiv preprint arXiv:1309.6352*, 2013.

[52] Saif M Mohammad and Peter D Turney. Nrc emotion lexicon. 2013.

[53] Clifford Nass and Kwan Min Lee. Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, 7(3):171, 2001.

[54] Maria Augusta SN Nunes and Rong Hu. Personality-based recommender systems: an overview. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 5–6. ACM, 2012.

[55] Jon Oberlander and Scott Nowson. Whose thumb is it anyway?: Classifying author personality from weblog text. In *Proc. of COLING/ACL*, pages 627–634, 2006.

[56] Ante Odic, Marko Tkalcic, Jurij F Tasic, and Andrej Košir. Relevant context in a movie recommender system: Users opinion vs. statistical detection. *ACM RecSys*, 12, 2012.

[57] Rodrigo De Oliveira, Mauro Cherubini, and Nuria Oliver. Influence of personality on satisfaction with mobile phone services. *TOCHI*, 20(2), 2013.

[58] Dean Peabody and Lewis R Goldberg. Some determinants of factor structures from personality-trait descriptors. *Journal of Personality and Social Psychology*, 57(3):552, 1989.

[59] James W Pennebaker, Cindy K Chung, Molly Ireland, Amy Gonzales, and Roger J Booth. The development and psychometric properties of liwc2007. *Austin, TX, LIWC. Net*, 2007.

[60] James W Pennebaker and Laura A King. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999.

[61] Robin L Plackett. Karl pearson and the chi-squared test. *International statistical review*, 51(1):59–72, 1983.

[62] Daniele Quercia, Michal Kosinski, David Stillwell, and Jon Crowcroft. Our Twitter profiles, our selves: Predicting personality with Twitter. In *Proceedings of SocialCom*, pages 180–185. IEEE, 2011.

[63] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014.

[64] Peter J Rentfrow and Samuel D Gosling. The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of personality and social psychology*, 84(6):1236, 2003.

[65] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press, 2002.

[66] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[67] Informatics Division Science and Oxon OX11 0QX Michael Wilson Engineering Research Council Rutherford Appleton Laboratory Chilton, Didcot. Mrc psycholinguistic database, April 1987.

[68] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[69] Yla R. Tausczik1 and James W. Pennebaker1. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[70] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *JASIST*, 61(12):2544–2558, 2010.

[71] Grigorios Tsoumakas, Eleftherios Spyromitros-Xioufis, Jozef Vilcek, and Ioannis Vlahavas. Mulan: A java library for multi-label learning. *The Journal of Machine Learning Research*, 12:2411–2414, 2011.

[72] Ernest C Tupes and Raymond E Christal. Recurrent personality factors based on trait ratings. Technical report, DTIC Document, 1961.

[73] Michael Wilson. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10, 1988.

[74] Eleftherios Spyromitros Xioufis, William Groves, Grigorios Tsoumakas, and Ioannis P. Vlahavas. Multi-label classification methods for multi-target regression. *CoRR*, abs/1211.6581, 2012.

Appendix A

CORRELATIONS FOR YOUTUBE VLOGGERS DATASET

The Spearman's rank correlation coefficients of all the extracted features with the 5 personality impression scores (*Extraversion (E)*, *Agreeableness (A)*, *Conscientiousness (C)*, *Emotional Stability (ES)*, *Openness (O)*) are presented in Table A.1–A.7. Significant ($p < 0.05$) correlations between features and personality impression scores are typeset in bold.

Table A.1: (Right) Correlation results between the emotional features and the 5 personality impression scores. (Left) Correlation results between the MRC features and the 5 personality impression scores.

NRC	E	A	C	ES	O
Positive	0.07	-0.01	**0.16**	0.08	0.02
Negative	0.07	**-0.29**	-0.05	**-0.19**	**-0.11**
Anger	0.10	**-0.29**	-0.06	**-0.16**	-0.05
Anticipation	0.02	-0.01	**0.13**	0.05	-0.06
Disgust	0.03	**-0.29**	-0.05	**-0.18**	-0.10
Fear	0.03	**-0.20**	0.03	**-0.14**	-0.08
Joy	0.09	0.07	**0.15**	0.10	0.03
Sadness	0.03	**-0.21**	0.03	**-0.14**	-0.06
Surprise	0.06	0.03	**0.12**	0.08	-0.05
Trust	0.03	-0.05	0.09	0.04	-0.02
SentiStrength	E	A	C	ES	O
Positive	0.04	**0.35**	**0.15**	**0.27**	0.10
Neutral	0.00	**-0.34**	**-0.19**	**-0.35**	**-0.20**
Negative	-0.03	**-0.12**	-0.05	-0.04	0.01

MRC	E	A	C	ES	O
NLET	-0.04	-0.09	**0.30**	0.07	0.00
NPHON	**-0.11**	-0.04	**0.17**	0.08	-0.05
NSYL	**-0.14**	-0.02	**0.16**	0.04	-0.07
KF FREQ	-0.03	-0.01	**0.25**	0.08	-0.02
KF NCATS	**-0.13**	0.02	0.07	-0.02	-0.02
KF NSAMP	**-0.17**	0.04	**0.15**	0.02	-0.05
TL FREQ	-0.03	0.01	**0.28**	0.09	-0.01
BROWN FREQ	-0.09	0.10	-0.06	-0.02	-0.07
FAM	-0.05	0.05	0.08	0.01	0.00
CONC	0.05	0.01	-0.06	-0.04	0.02
IMAG	0.08	0.03	-0.04	-0.04	0.03
MEANC	0.01	0.04	**-0.11**	-0.08	-0.01
MEANP	**0.15**	0.05	0.10	0.05	0.07
AOA	0.06	-0.07	**0.18**	0.09	0.00

Table A.2: Correlation results between gender and the 5 personality impression scores. The absolute value of Spearman's rank is reported here.

Feature	E	A	C	ES	O
gender	.02	**0.21**	.02	.02	.07

Table A.3: Correlation results between the LIWC features and the 5 personality impression scores. (Right) 1-25 Features. (Left) 26-51 Features.

LIWC Features	E	A	C	ES	O
Word Count	-0.07	**0.10**	**-0.15**	0.01	0.05
Words Per Sentence	**0.13**	-0.06	**-0.26**	**-0.11**	0.01
Sixltr	-0.01	0.05	**-0.27**	-0.07	-0.03
Dictionary words	**0.22**	-0.08	0.01	0.02	**0.10**
Numerals	0.02	-0.05	-0.03	-0.07	**0.10**
funct	**0.18**	-0.07	-0.08	-0.03	0.07
Pronouns	0.04	-0.08	**0.15**	0.03	0.07
ppron	-0.05	-0.08	**0.18**	0.08	0.02
i	0.03	**-0.12**	**0.23**	0.06	0.03
we	-0.07	0.03	**-0.10**	-0.04	-0.02
you	**-0.14**	-0.02	-0.02	-0.02	-0.07
shehe	-0.09	**0.10**	0.01	0.07	-0.03
they	0.06	**0.14**	**-0.11**	0.03	0.03
ipron	**0.15**	0.00	-0.03	-0.08	0.04
article	0.00	0.08	**-0.12**	-0.09	-0.02
verb	0.08	-0.02	**0.19**	0.07	0.08
auxverb	0.06	0.00	**0.18**	0.08	0.06
past	0.08	-0.02	0.06	0.04	0.05
present	-0.02	0.01	**0.17**	0.07	0.04
future	0.09	-0.02	0.01	-0.04	0.08
adverb	0.05	**-0.11**	**0.13**	-0.03	0.06
preps	0.06	0.01	**-0.28**	-0.07	-0.08
conj	0.08	**-0.16**	-0.05	-0.02	0.04
negate	0.03	**0.17**	**0.22**	0.18	0.05
quant	0.09	**0.15**	**-0.10**	**0.02**	0.06

LIWC Features	E	A	C	ES	O
number	-0.05	0.03	-0.06	-0.05	0.03
swear	-0.08	**0.39**	**0.20**	**0.25**	0.04
social	-0.07	0.05	**-0.16**	0.00	-0.03
family	-0.07	0.02	-0.07	0.05	-0.04
friend	0.00	-0.09	-0.01	0.00	**-0.12**
humans	**-0.12**	0.08	-0.02	0.02	-0.06
affect	**-0.12**	0.00	0.06	0.06	-0.04
posemo	-0.09	**-0.27**	-0.08	**-0.15**	**-0.12**
negemo	-0.03	**0.40**	**0.27**	**0.35**	**0.14**
anx	-0.06	0.04	-0.01	0.07	-0.02
anger	-0.07	**0.41**	**0.27**	**0.27**	0.07
sad	0.06	0.00	0.03	0.03	**0.13**
cogmech	**0.20**	-0.09	-0.09	-0.02	0.08
insight	0.04	0.02	-0.06	-0.02	0.01
cause	0.08	-0.02	0.07	0.03	0.05
discrep	**0.12**	-0.08	**-0.11**	**-0.11**	-0.02
tentat	**0.21**	0.05	-0.02	-0.08	0.07
certain	-0.01	0.07	-0.04	0.07	0.02
inhib	0.09	0.09	-0.01	0.08	0.07
incl	-0.04	**-0.14**	**-0.23**	-0.01	-0.03
excl	**0.14**	-0.06	0.08	0.00	**0.10**
percept	-0.06	-0.02	0.02	-0.05	-0.06
see	-0.08	0.01	0.01	-0.03	-0.03
hear	-0.01	0.09	0.00	0.00	-0.08
feel	0.05	-0.09	0.03	0.03	0.05
bio	-0.05	**0.12**	**0.12**	**0.19**	0.08

Table A.4: Correlation results between the LIWC features and the 5 personality impression scores. 52-81 Features.

LIWC Features	E	A	C	ES	O
body	0.00	**0.15**	**0.18**	**0.15**	0.08
health	0.03	0.08	-0.02	**0.11**	**0.10**
sexual	**-0.19**	**0.17**	**0.13**	**0.22**	-0.02
ingest	-0.04	-0.04	-0.07	-0.04	-0.03
relativ	-0.08	0.00	-0.01	0.07	-0.05
motion	-0.06	0.02	0.03	0.10	**-0.11**
space	**-0.15**	0.03	**-0.10**	-0.01	-0.08
time	0.05	-0.03	**0.15**	0.08	0.07
work	0.08	**-0.13**	**-0.24**	**-0.13**	-0.05
achieve	0.02	-0.06	**-0.20**	**-0.10**	0.07
leisure	-0.06	-0.06	0.01	**-0.12**	**-0.14**
home	-0.02	-0.01	-0.02	0.03	0.04
money	-0.05	-0.02	-0.10	-0.07	-0.02
relig	**-0.13**	**0.11**	0.05	0.03	-0.02
death	0.04	**0.16**	0.00	0.10	0.06
assent	**-0.17**	-0.03	**0.26**	0.05	-0.08
nonfl	**0.18**	**-0.13**	0.06	-0.09	**0.14**
filler	-0.03	-0.03	**0.23**	0.07	0.05
Period	-0.05	0.01	**0.22**	0.08	0.00
Comma	0.03	-0.01	**0.14**	-0.02	0.04
Colon	0.00	0.04	-0.08	0.01	**0.10**
SemiC	-0.06	0.07	0.05	0.03	0.03
QMark	**-0.22**	**0.19**	**0.19**	**0.15**	-0.03
Exclam	**-0.21**	-0.01	0.08	0.00	**-0.13**
Quote	0.01	0.02	-0.01	0.04	-0.02
Dash	**0.10**	**0.10**	**0.15**	0.07	**0.11**
Apostro	0.00	0.02	**0.26**	0.08	0.07
Parenth	-0.02	0.03	-0.01	0.08	0.00
OtherP	-0.04	-0.06	-0.07	**-0.10**	-0.09
AllPct	0.00	0.07	**0.31**	0.08	0.06

Table A.5: Correlation results between the audio-video features and the 5 personality impression scores.

Speaking Activity	E	A	C	ES	O
Time speaking	**0.18**	0.07	**0.27**	**0.15**	**0.14**
Avg length segm	**0.17**	0.04	**0.17**	**0.11**	**0.12**
No. of turns	**-0.17**	0.01	-0.04	-0.04	-0.09
Prosodic Cues	E	A	C	ES	O
mean pitch	**0.23**	0.09	-0.08	-0.06	0.07
sd pitch	**-0.12**	**-0.17**	0.01	-0.02	-0.02
meanconf pitch	**0.21**	**0.14**	0.00	0.01	0.04
sdconf pitch	**0.16**	**0.11**	0.03	0.02	0.06
mean spec entropy	**0.11**	-0.05	-0.10	0.00	-0.02
sd spec entropy	0.02	0.01	0.02	-0.05	0.02
mean val apeak	-0.03	0.10	0.01	0.01	-0.04
sd val apeak	0.00	-0.06	0.01	0.01	0.04
mean loc apeak	**0.29**	0.02	-0.06	-0.08	0.05
sd loc apeak	-0.03	**-0.13**	-0.10	-0.08	-0.06
mean num apeak	**0.17**	-0.05	-0.08	-0.06	-0.02
sd num apeak	0.06	-0.10	-0.10	-0.07	-0.08
mean energy	**0.24**	-0.10	-0.08	-0.03	0.04
sd energy	0.10	-0.04	-0.09	-0.04	0.09
mean d energy	-0.06	0.02	0.03	0.02	-0.06
sd d energy	**0.31**	**-0.12**	**-0.11**	-0.05	0.08
avg voiced seg	-0.05	**-0.12**	-0.07	-0.06	-0.08
voice rate	0.01	**0.11**	0.04	0.04	0.05
Visual Activity	E	A	C	ES	O
hogv entropy	**0.32**	-0.04	**-0.22**	-0.08	**0.19**
hogv median	**0.29**	0.03	**-0.16**	-0.02	**0.22**
hogv cogR	0.01	-0.06	-0.02	0.05	0.01
hogv cogC	0.00	-0.01	-0.04	-0.06	-0.06

Table A.6: Correlation results between the Splice features and the 5 personality impression scores. 1-37 Features.

Splice Features	E	A	C	ES	O
numChars	0.08	-0.10	**0.12**	0.00	-0.03
numCharsMinusSpacesAndPunctuation	0.08	-0.10	**0.13**	0.01	-0.03
numWords	0.08	-0.10	**0.11**	0.00	-0.03
numSentences	**0.13**	**-0.14**	0.00	-0.05	-0.04
numPunctuation	0.10	**-0.12**	-0.01	-0.03	-0.04
numNouns	0.10	-0.09	0.09	0.00	-0.03
nounRatio	**0.13**	0.03	**-0.16**	0.00	0.02
numVerbs	0.07	-0.09	0.10	-0.01	-0.03
verbRatio	-0.04	-0.01	-0.04	-0.08	-.02
numAdjectives	0.10	**-0.11**	0.10	-0.01	-0.01
adjectiveRatio	0.10	-0.05	0.00	0.00	.07
numAdverbs	0.05	-0.08	0.09	-0.01	-.03
adverbRatio	-0.06	0.03	-0.05	-0.04	0.00
firstPersonSingular	0.06	-0.01	-0.04	-0.03	-.01
firstPersonPlural	0.07	-0.05	**0.14**	0.04	-0.01
secondPerson	**0.14**	-0.04	0.10	0.02	0.00
thirdPersonSingular	-0.02	**-0.18**	0.10	-0.01	-0.05
thirdPersonPlural	-0.02	**-0.18**	0.10	-0.01	-0.05
iCanDoIt	0.01	-0.06	-0.01	-0.01	-0.05
doKnow	0.06	0.01	-0.03	-0.01	-0.07
posSelfImage	0.07	0.03	-0.03	0.02	0.02
iCantDoIt	-0.02	-0.06	-0.07	-0.04	-0.07
dontKnow	0.03	-0.02	-0.09	-0.02	0.04
negSelfImage	0.01	**-0.11**	-0.09	-0.07	0.01
numImperatives	**0.16**	-0.03	-0.03	0.00	-0.01
suggestionPhrases	-0.02	-0.02	**0.14**	0.05	-0.06
inflexibility	0.01	-0.02	0.02	-0.06	-0.10
contradict	0.03	-0.05	0.01	0.02	0.00
totalDominance	**0.11**	-0.04	0.04	0.01	-0.06
dominanceRatio	0.03	0.10	0.06	0.10	-0.03
numAgreement	0.09	-0.07	-0.05	0.01	-0.03
agreementRatio	0.05	-0.01	**-0.16**	0.02	-0.01
totalSubmissiveness	0.04	-0.10	**-0.14**	-0.07	-0.02
submissivenessRatio	-0.01	-0.04	**-0.14**	-0.04	-0.01
Imagery	**0.19**	-0.02	0.07	0.01	**0.11**
Pleasantness	**0.13**	**0.26**	0.05	**0.20**	**0.17**
Activation	**0.11**	-0.08	-0.01	-0.09	0.06

Table A.7: Correlation results between the Splice features and the 5 personality impression scores. 38-74 Features.

Splice Features	E	A	C	ES	O
avgWordLength	0.02	-0.05	**0.32**	**0.12**	0.03
avgSentenceLength	**-0.14**	0.09	**0.23**	**0.14**	-0.03
numSyllables	0.08	-0.10	**0.12**	0.01	-0.03
avgSyllablesPerWord	-0.01	0.03	**0.28**	**0.14**	-0.01
numWords3OrMoreSyll	0.07	**-0.11**	**0.18**	0.03	-0.02
rateWords3OrMoreSyll	0.00	-0.03	**0.23**	**0.12**	0.02
numWords6OrMoreChars	0.06	**-0.11**	**0.15**	0.01	-0.03
rateWords6OrMoreChars	-0.04	-0.04	0.**22**	0.07	0.01
numWords7OrMoreChars	0.07	-0.10	**0.18**	0.02	-0.02
rateWords7OrMoreChars	0.01	0.01	**0.29**	**0.12**	0.06
LexicalDiversity	-0.05	0.06	-0.06	0.01	0.05
hedgeVerb	-0.01	-0.03	0.10	0.04	-0.06
hedgeConj	-0.02	-0.07	0.06	0.01	-0.04
hedgeAdj	-0.05	-0.10	0.08	0.00	-0.08
hedgeModal	-0.05	-0.03	**0.16**	0.04	0.00
hedgeAll	-0.04	-0.10	**0.11**	0.01	-0.04
numDisfluencies	-0.01	0.01	-0.05	0.06	-0.06
disfluencyRatio	-0.07	0.10	**-0.12**	0.08	-0.07
numInterjections	**0.11**	**-0.11**	**-0.11**	-0.09	-0.03
interjectionRatio	0.06	-0.08	**-0.28**	**-0.13**	-0.03
numSpeculate	-0.03	**-0.16**	0.02	0.00	-0.08
Expressivity	**-0.11**	0.09	-0.09	0.00	0.03
Pausality	**0.13**	-0.07	0.02	-0.06	0.06
questionCount	**0.22**	**-0.22**	**-0.11**	**-0.14**	-0.01
questionRatio	**0.20**	**-0.21**	**-0.15**	**-0.15**	0.01
pastTense	0.02	-0.03	0.08	-0.01	-0.05
presentTense	0.09	**-0.11**	0.07	-0.03	-0.03
ARI	-0.10	0.06	**0.30**	**0.14**	-0.01
FRE	0.09	-0.07	**-0.31**	**-0.16**	0.02
CLI	**0.14**	-0.09	**-0.23**	**-0.14**	0.03
LWRF	**-0.13**	0.09	**0.25**	**0.15**	-0.02
FOG	**-0.11**	0.07	**0.28**	**0.16**	-0.01
SMOG	-0.07	0.04	**0.29**	**0.16**	0.00
DALE	**-0.14**	0.09	**0.23**	**0.14**	-0.02
LIX	**-0.12**	0.06	**0.29**	**0.13**	0.00
RIX	-0.10	0.08	**0.32**	**0.17**	0.01
FRY	**-0.13**	0.09	**0.26**	**0.15**	-0.02

Appendix B

CORRELATIONS FOR MYPERSONALITY DATASET

The Pearson's correlation coefficients of all the extracted features with the 5 personality scores (*Extraversion (E)*, *Agreeableness (A)*, *Conscientiousness (C)*, *Neuroticm (Neu)*, *Openness (O)*) are presented in Table B.1–B.5. Significant ($p < 0.05$) correlations between features and personality scores are typeset in bold.

Table B.1: Correlation results between the Demographic and Facebook Activity features and the 5 personality scores.

Demographic/ FB Activity	E	A	C	Neu	O
age	-0.01	**0.05**	**0.17**	**-0.04**	0.00
gender	**0.02**	**0.06**	**0.05**	**0.18**	**0.023**
network size	**0.20**	**0.04**	**0.04**	**-0.07**	-0.00
#like	0.00	**-0.03**	**-0.09**	**0.08**	**0.02**
#group	**0.06**	**-0.01**	**-0.06**	**0.03**	**0.05**
#education	0.01	**0.02**	**0.09**	**-0.04**	**0.05**
#diad	**0.13**	**0.02**	**0.03**	**-0.04**	**0.02**
#status	**0.07**	**-0.01**	**-0.03**	**0.07**	**0.05**
SentiStrength	E	A	C	ES	O
Positive	0.043	0.043	0.027	0.001	-0.004
Negative	0.004	0.011	0.021	-0.036	-0.019

MRC	E	A	C	ES	O
NLET	**-0.03**	**0.05**	**0.08**	-0.01	**0.07**
NPHON	**-0.05**	**0.03**	**0.05**	-0.00	**0.08**
NSYL	**-0.03**	**0.05**	**0.08**	0.00	**0.07**
KF FREQ	**-0.01**	**0.05**	**0.09**	**-0.04**	**0.08**
KF NCATS	-0.00	**0.06**	**0.08**	0.00	**0.06**
KF NSAMP	0.00	**0.06**	**0.08**	**-0.01**	**0.06**
TL FREQ	-0.01	**0.05**	**0.09**	**-0.04**	**0.08**
BROWN FREQ	-0.01	**0.05**	**0.06**	-0.01	**0.10**
FAM	0.00	**0.06**	**0.07**	-0.00	**0.06**
CONC	0.01	**0.05**	**0.06**	-0.00	**0.05**
IMAG	0.01	**0.06**	**0.06**	-0.00	**0.05**
MEANC	**0.01**	**0.05**	**0.06**	0.003	**0.05**
MEANP	**0.02**	**0.04**	**0.05**	**-0.03**	**0.02**
AOA	-0.01	0.01	**0.04**	**-0.03**	**0.04**

Table B.2: Correlation results between the LIWC features and the 5 personality impression scores. (Right) 1-25 Features. (Left) 26-51 Features.

LIWC Features	E	A	C	ES	O
WC	**0.07**	**0.02**	**0.01**	**0.01**	**0.05**
WPS	**0.02**	**-0.03**	**-0.02**	-0.00	**-0.05**
Sixltr	**-0.03**	**-0.02**	-0.01	**-0.01**	0.01
Dic	0.01	**0.06**	**0.06**	0.01	**0.05**
Numerals	0.00	-0.01	-0.01	**-0.01**	**-0.05**
funct	-0.01	**0.05**	**0.05**	0.01	**0.07**
pronoun	0.01	0.01	-0.01	**0.04**	**0.06**
ppron	**0.02**	0.00	**-0.02**	**0.05**	**0.05**
i	**0.02**	0.00	**-0.04**	**0.04**	**0.04**
we	**0.02**	**0.02**	**0.04**	**-0.02**	0.00
you	0.01	0.00	0.00	0.01	**0.03**
shehe	0.00	-0.00	0.01	**0.04**	**0.03**
they	**-0.02**	**-0.01**	**0.01**	0.00	**0.02**
ipron	**-0.02**	**0.01**	**0.02**	0.00	**0.05**
article	**-0.02**	**0.03**	**0.04**	**-0.05**	**0.07**
verb	0.00	**0.03**	-0.00	**0.04**	0.02
auxverb	**-0.01**	**0.02**	-0.00	**0.03**	**0.04**
past	**-0.01**	**0.02**	-0.01	0.01	-0.00
present	**0.01**	**0.01**	-0.01	**0.04**	**0.01**
future	-0.01	0.00	0.00	0.01	**0.04**
adverb	-0.01	**0.02**	**-0.01**	**0.05**	**0.02**
preps	0.00	**0.05**	**0.09**	**-0.03**	**0.02**
conj	**0.01**	**0.02**	**0.02**	**0.03**	**0.04**
negate	**-0.02**	**-0.01**	**-0.03**	**0.04**	**0.03**
quant	-0.01	**0.03**	**0.04**	**-0.02**	0.01

LIWC Features	E	A	C	ES	O
number	**-0.02**	0.01	**0.01**	**-0.01**	**0.03**
swear	**0.02**	**-0.09**	**-0.08**	**0.03**	0.01
social	**0.03**	**0.01**	**0.02**	**0.01**	**0.01**
family	0.01	**0.03**	**0.04**	**0.01**	**-0.05**
friend	**0.02**	**0.02**	**0.02**	0.00	**-0.01**
humans	**0.03**	**-0.02**	0.00	0.00	0.00
affect	**0.04**	**0.01**	0.00	**0.03**	**-0.03**
posemo	**0.06**	**0.08**	**0.08**	**-0.02**	**-0.04**
negemo	**-0.02**	**-0.08**	**-0.10**	**0.08**	**0.01**
anx	**-0.02**	0.00	**-0.02**	**0.03**	**0.02**
anger	**-0.01**	**-0.10**	**-0.09**	**0.05**	**0.01**
sad	-0.01	-0.01	**-0.02**	**0.05**	**0.01**
cogmech	**-0.01**	**0.02**	**0.02**	**0.03**	**0.07**
insight	**-0.03**	0.00	-0.01	**0.03**	**0.06**
cause	**-0.03**	-0.01	**-0.01**	**0.01**	**0.04**
discrep	**-0.01**	0.01	0.00	**0.02**	**0.01**
tentat	**-0.02**	0.00	-0.01	**0.01**	**0.04**
certain	0.00	0.01	**0.02**	0.00	**0.02**
inhib	0.00	0.01	**0.02**	**0.01**	0.00
incl	**0.04**	**0.05**	**0.06**	0.00	**0.04**
excl	**-0.02**	0.00	**-0.02**	**0.03**	**0.03**
percept	**-0.02**	0.00	**-0.03**	**0.02**	**0.05**
see	-0.01	0.01	-0.01	0.01	**0.02**
hear	**-0.02**	-0.01	**-0.04**	0.01	**0.05**
feel	0.00	0.00	**-0.02**	**0.02**	**0.01**
bio	**0.03**	**-0.03**	**-0.05**	**0.04**	**0.02**

Table B.3: Correlation results between the LIWC features and the 5 personality impression scores. 52-81 Features.

LIWC Features	E	A	C	ES	O
body	0.01	**-0.04**	**-0.06**	**0.03**	**0.03**
health	0.01	0.00	0.00	**0.04**	0.00
sexual	**0.05**	**-0.03**	**-0.04**	**0.02**	0.00
ingest	0.00	0.00	-0.01	0.00	**0.02**
relativ	**0.02**	**0.07**	**0.09**	**-0.04**	**-0.03**
motion	**0.02**	**0.04**	**0.04**	**-0.01**	**-0.02**
space	0.01	**0.03**	**0.03**	**-0.03**	**0.04**
time	**0.01**	**0.06**	**0.09**	**-0.02**	**-0.07**
work	**-0.04**	**0.02**	**0.05**	**-0.02**	**-0.03**
achieve	-0.01	**0.02**	**0.05**	**-0.03**	-0.02
leisure	**0.04**	**0.03**	**0.03**	**-0.04**	-0.01
home	0**.01**	**0.03**	**0.03**	0.00	**-0.03**
money	0.00	-0.01	**0.02**	**-0.01**	**0.02**
relig	-0.01	**0.03**	**0.03**	**-0.02**	0.01
death	**-0.03**	**-0.04**	**-0.04**	**0.02**	**0.05**
assent	**0.03**	**0.02**	**-0.04**	0.00	**-0.02**
nonfl	0.00	0.00	-0.01	**0.01**	0.00
filler	0.00	**-0.01**	**-0.03**	0.01	**0.02**
Period	0.00	-0.01	0.00	0.00	**-0.03**
Comma	-0.01	**-0.02**	**-0.03**	0.01	**-0.01**
Colon	0.01	**0.01**	**-0.03**	**0.02**	**-0.06**
SemiC	-0.01	-0.01	-0.01	0.01	0.00
Qmark	0.00	-0.01	0.00	0.00	0.00
Exclam	**0.03**	**0.01**	0.00	0.00	**-0.04**
Dash	**-0.03**	**-0.02**	-0.01	0.00	0.01
Quote	**-0.01**	0.00	-0.01	0.00	**-0.01**
Apostro	**-0.06**	**-0.01**	**-0.03**	**0.05**	**0.08**
Parenth	**0.02**	0.01	**-0.02**	**0.01**	**-0.04**
OtherP	**-0.02**	**-0.01**	**-0.01**	0.00	0.00
AllPct	0.00	-0.01	0.00	0.00	0.00

Table B.4: Correlation results between the Splice features and the 5 personality impression scores. 1-32 Features.

Splice Features	E	A	C	ES	O
numChars	**0.07**	**0.01**	-0.01	**0.01**	**0.05**
numCharsMinus Spaces&Punctuation	**0.06**	0.01	0.00	**0.01**	**0.07**
numWords	**0.07**	**0.02**	**0.01**	**0.01**	**0.05**
numSentences	**0.03**	**0.04**	**0.02**	**0.01**	**0.03**
numPunctuation	**0.02**	0.00	**-0.02**	0.01	**-0.04**
numNouns	**0.08**	**-0.02**	**-0.05**	**0.01**	**-0.03**
numVerbs	**0.01**	**0.03**	**0.02**	**0.03**	**0.06**
numAdjectives	**0.02**	**0.04**	**0.03**	0.00	**0.04**
numAdverbs	**0.01**	**0.01**	0.00	**0.03**	**0.03**
firstPersonSingular	**0.05**	0.00	**-0.04**	**0.04**	**0.02**
firstPersonPlural	**0.03**	**0.03**	**0.04**	**-0.02**	0.00
secondPerson	**0.02**	0.00	0.01	0.00	**0.04**
thirdPersonSingular	**-0.02**	**-0.01**	**0.01**	0.00	**0.02**
thirdPersonPlural	**-0.02**	**-0.01**	**0.01**	0.00	**0.02**
iCanDoIt	0.00	0.00	0.00	0.01	**0.02**
doKnow	0.00	0.00	0.00	0.01	**0.02**
posSelfImage	**0.01**	**0.02**	**0.02**	0.00	0.01
iCantDoIt	**-0.01**	0.01	0.00	**0.02**	**0.03**
dontKnow	**-0.02**	0.00	**-0.01**	**0.02**	**0.02**
negSelfImage	-0.01	0.01	0.01	**0.02**	**0.01**
numImperatives	0.01	**0.01**	**0.01**	0.00	**0.02**
suggestionPhrases	0.01	-0.01	**-0.02**	0.01	**0.03**
inflexibility	**-0.02**	-0.01	**-0.01**	0.01	0.01
contradict	0.00	0.00	0.00	0.01	0.01
totalDominance	0.01	**0.01**	0.01	0.01	**0.04**
numAgreement	**0.03**	**0.02**	0.01	-0.01	0.00
askPermission	0.00	0.00	-0.01	0.00	0.00
seekGuidance	0.01	**0.01**	-0.01	0.00	0.01
totalSubmissiveness	**-0.02**	**0.01**	0.00	**0.03**	**0.04**
Imagery	**0.02**	-0.01	**-0.03**	0.01	**-0.02**
Pleasantness	**0.05**	**0.04**	**0.01**	-0.01	**-0.02**
Activation	**0.03**	0.01	**-0.02**	0.01	**-0.02**

Table B.5: Correlation results between the Splice features and the 5 personality impression scores. 33-71 Features.

Splice Features	E	A	C	ES	O
avgWordLength	**-0.02**	**-0.01**	-0.01	0.00	0.01
avgSentenceLength	0.01	**-0.02**	**-0.03**	0.00	**-0.04**
numSyllables	**0.06**	0.01	0.00	0.01	**0.08**
avgSyllablesPerWord	**-0.03**	-0.02	0.00	-0.01	**0.02**
numWordsWith3OrMoreSyllables	0.00	-0.02	0.00	**-0.02**	**0.06**
rateWordsWith3OrMoreSyllables	**-0.03**	**-0.03**	0.00	**-0.02**	**0.02**
numWordsWith6OrMoreChars	**0.03**	0.00	**-0.01**	**0.02**	**0.06**
rateWordsWith6OrMoreChars	**-0.04**	**-0.02**	**-0.03**	0.01	**0.01**
numWordsWith7OrMoreChars	**0.03**	0.01	0.00	0.00	**0.05**
rateWordsWith7OrMoreChars	**-0.03**	**-0.02**	**-0.01**	0.00	0.00
LexicalDiversity	**-0.03**	**-0.04**	**-0.05**	0.00	0.00
hedgeVerb	0.00	0.00	-0.01	**0.03**	**0.04**
hedgeConj	0.00	-0.01	**-0.01**	0.01	**0.03**
hedgeAdj	0.00	**0.01**	**0.01**	-0.01	**0.04**
hedgeModal	**-0.02**	0.00	0.00	**0.02**	**0.04**
hedgeAll	-0.01	0.01	0.00	**0.02**	**0.07**
numDisfluencies	**0.02**	**-0.01**	**-0.04**	**0.01**	0.00
numInterjections	**0.03**	0.00	**-0.04**	**0.02**	**0.02**
numSpeculate	**-0.01**	0.00	0.00	**0.02**	**0.04**
Expressivity	**-0.05**	0.01	**0.02**	0.00	-0.01
Pausality	**-0.04**	**0.03**	**0.05**	0.00	**0.09**
questionCount	**-0.02**	**-0.02**	**-0.02**	0.01	0.00
pastTense	-0.01	**0.02**	0.00	**0.01**	**0.03**
presentTense	0.01	**0.03**	**0.02**	**0.03**	**0.07**
ARI	-0.01	**-0.02**	**-0.03**	0.00	**-0.03**
FRE	**0.02**	**0.03**	**0.02**	0.01	0.00
CLI	-0.01	**0.02**	**0.03**	0.00	**0.04**
LWRF	0.01	**-0.02**	**-0.03**	0.00	**-0.04**
FOG	0.00	**-0.02**	**-0.03**	-0.01	**-0.03**
SMOG	0.00	**-0.03**	**-0.02**	**-0.02**	-0.01
DALE	0.01	**-0.02**	**-0.03**	0.00	**-0.04**
LIX	-0.01	**-0.02**	**-0.04**	0.00	**-0.03**
RIX	0.00	**-0.02**	**-0.03**	-0.01	**-0.04**
FRY	0.01	**-0.02**	**-0.03**	-0.01	**-0.03**
numPassiveVerbs	**-0.02**	**0.01**	**0.03**	**0.01**	**0.03**
SWNpositivity	**0.01**	**0.05**	**0.05**	0.00	**0.03**
SWNnegativity	**-0.02**	**-0.02**	**-0.02**	**0.06**	**0.05**
SWNobjectivity	**0.02**	-0.01	-0.01	**-0.01**	**-0.01**
FKGL	-0.01	**-0.03**	**-0.03**	0.00	**-0.02**

Appendix C

CORRELATIONS BETWEEN EMOTIONS AND FACEBOOK FEATURES

To assess the relation between the different features and emotions, we apply the Pearson chi-squared dependence test [61]. Table C.1 presents the p-values. The null hypothesis is that features and emotions are independent. The p-values that are lower than the significance level ($p < .01$) denote significant correlations of features with emotions. They are indicated in bold in the table. Gender is related with all emotion categories except fear. Age is shown to be related to all emotion types. Similarly, Openness and Extroversion are related with all emotion types. Conscientiousness is related to anger, negative and sadness emotions. Agreeableness is not related to sadness. And finally, Neuroticism shows no relation with positive, anticipation and trust emotions.

Table C.1: Pearson Chi Squared test results for on characteristics of users and posts, and emotion categories: *positive (Pos)*, *negative (Neg)*, *anger (Ang)*, *anticipation (Ant)*, *disgust (Dis)*, *fear (Fea)*, *joy (Joy)*, *sadness (Sad)*, *surprise (Sur)*, and *trust (Tru)*.

Features	Pos	Neg	Ang	Ant	Dis	Fea	Joy	Sad	Sur	Tru
Gender	**0**	**0**	**0**	**0**	**0**	0.205	**0**	**0**	**0**	**0**
Age	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Open	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Conscientious	**0**	0.014	0.793	**0**	**0**	**0**	**0**	0.496	**0**	**0**
Extrovert	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Agreeable	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0.065	**0**	**0**
Neurotic	0.613	**0**	**0**	0.015	**0**	**0**	**0**	**0**	**0**	0.249
Monday	**0.001**	0.023	0.146	0.050	0.058	0.333	**0**	0.105	0.055	0.081
Tuesday	**0.001**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	0.029	0.879
Wednesday	0.213	**0**	**0**	0.137	**0**	**0**	**0.001**	**0**	0.220	**0.001**
Thursday	**0**	**0**	**0.008**	**0.002**	**0**	**0**	**0.002**	**0.001**	0.482	**0**
Friday	0.019	0.029	0.517	**0**	0.139	0.566	**0.001**	0.047	**0**	**0**
Saturday	0.437	**0**	**0**	0.891	**0**	**0**	**0**	**0**	0.104	**0**
Sunday	0.029	**0**	**0**	0.170	**0**	**0**	**0**	**0**	0.200	**0**
January	0.039	**0**	0.016	0.704	**0.007**	**0.003**	**0**	**0**	0.066	**0.004**
February	0.432	0.377	0.740	**0.001**	0.035	0.322	**0.006**	0.082	0.094	0.139
March	**0**	**0**	**0.001**	0.019	**0**	**0**	**0.004**	**0**	**0**	**0**
April	**0.003**	0.027	**0.002**	**0.002**	**0.001**	**0.002**	**0.004**	0.025	**0.002**	**0**
May	**0.001**	0.016	0.465	0.269	0.623	**0.004**	**0**	**0.003**	**0.003**	**0**
June	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
July	**0**	**0**	0.011	**0.001**	**0**	**0**	**0**	**0**	**0**	**0**
August	**0**	**0**	**0.001**	0.424	**0**	**0**	**0**	**0**	**0**	**0.001**
September	**0**	0.014	**0**	**0.004**	**0.003**	0.016	**0**	**0.008**	**0**	0.027
October	**0**	**0**	0.614	**0**	**0.001**	**0.003**	**0**	**0.009**	**0.004**	**0**
November	**0.001**	0.130	0.126	0.760	**0.002**	0.410	**0**	**0.003**	0.250	0.513
December	**0**	0.113	**0.004**	**0.004**	0.004	0.211	**0**	0.195	**0**	**0**