Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

A Study of Politeness Theories on Social Media Forums

by Krishna Hariramani

Presented to the University of Dublin, Trinity College
in partial fulfilment of the requirements for the degree of
Master of Science in Computer Science - Data Science

Supervisor: Prof. Carl Vogel

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Krishna Hariramani

August 14, 2019

Acknowledgments

Firstly, I would like to thank my dissertation advisor Professor Carl Vogel of the School of Computer Science and Statistics, Trinity College Dublin. He played his role as a mentor superbly throughout my year of study and during the period of the dissertation. He guided me in every aspect of the dissertation, from formulating the research question to steering me in the right direction whenever needed with his advice. I would also like to thank my parents and sister from whom I've been living away for more than a year now to complete my Master's degree for their continuous support and encouragement. I would also like to thank my friends who never failed to encourage me and were always available for me whenever I needed them the most. I would also like to express my gratitude to all my Professors who have taught me the tools and techniques to perform this research. Finally, I would like to thank Trinity College Dublin for providing me the opportunity to perform this research. Thank you.

Abstract

This study investigates the use of popular politeness theories over social media forums like Stack Exchange and Reddit. Various popular politeness theories are hypothesized and tested over social media forums. In order to conduct the hypothesis testing, thorough investigation of various tools, techniques, and datasets for predicting politeness scores of textual data is conducted. Along with data pre-processing of investigated datasets, data collection over these social media forums is carried out to collect new datasets so that we have datasets befitting our hypotheses. Investigated tools and techniques are then used for training machine learning models and conduct politeness prediction over our datasets. We perform rigorous hypothesis testing using F-test, T-tests and Person correlation test. Results of this study discover very interesting insights into how politeness theories work on social media forums. One interesting insight from this research is that comments posted by females are more polite when compared to comments posted by males, however, comments posted by females don't get more polite replies when compared to replies made to comments posted by males, nevertheless, replies made to comments posted by users of the opposite gender are more polite than replies made to comments posted by users of the same gender.

Keywords: Politeness theory, Computational linguistics, Social Media, Machine Learning, LASSO regression, Data Analytics

Contents

List of Figures

List of Tables

Abbreviations

TV	Television
AMT	Amazon Mechanical Turk
BOW	Bag of Words
CNN	Convolutional Neural Networks
SVM	Support Vector Machine
CRAN	The Comprehensive R Archive Network
API	Application Program Interface
CSV	Comma Separated values
PRAW	Python Reddit Application Program Interface Wrapper
FTP	File Transfer Protocol
LASSO	Least Absolute Shrinkage and Selection Operator

Chapter 1

Introduction

Politeness theory has been a topic of high interest for the researchers. Most influential works in this field are works of Brown and Levison [4], Lakkoff [16] and Culpeper[6, 7]. According to Brown and Levinson, "politeness is defined as strategies used by speakers to bring back equilibrium during a conversation after a face-threatening act is committed". With the emergence of the internet, huge datasets are now available for research. Also, because of the high availability of data, the study of machine learning techniques has been on a rise recently. In the modern-day era, social media forums provide a place where people can share their opinions on various topics and interact with others based on their opinions. The advent of the internet and social media, and recent advancements in machine learning research provide great opportunities for discourse studies.

This study tests a few hypotheses based on popular politeness theories over social media forums such as Reddit and Stack Exchange. A thorough background investigation is conducted on past research to get a fair idea of the already present literature and find out gaps in the research. The investigated literature is then used to determine the relevant tools, techniques, and datasets required for hypothesis testing. Along with data pre-processing on investigated datasets, data collection over Reddit using the python PRAW library [3] is also performed so that our hypotheses can be tested on befitting datasets. The CRAN politeness package [5] is then used to train a LASSO regression model over existing pre-processed manually annotated dataset to predict politeness scores of the collected dataset. Pearson's r correlation test is then used to test hypotheses which try to correlate politeness scores with the variables of social power on social media (reputation, upvotes, downvotes, and user karma). F-test and T-test are also used to test the hypotheses which try to find a relation between politeness scores and gender of the users on Reddit. Finally, all the results are analyzed to get interesting insights into how politeness theory works on social media. The room for future work is also outlined.

Section 1.1 lays out the motivation to carry out this study, Section 1.2 states the research question this study tries to answer, Section 1.3 states the research aim of this study, Section 1.5 explains the

overall structure of the thesis.

1.1 Motivation

Brown and Levinson state in their work that, “politeness is a universal dimension of human communication” and “politeness is a universal construct” [4]. Researchers in many domains might want to compare this universal construct (politeness) and markers related with it to other covariates of interests like social power, gender, etc. For example, researchers might want to know whether, after an experiment, the subject is speaking more politely or not. Or some researchers might want to compare how responses of people vary to polite and impolite language. There is also a good amount of research going on the subject of how interventions in speech could affect politeness and whether politeness is correlated to some other covariates which are situational or trait-level. There could also be a possibility of generating a model based on the coalescing of these analytical approaches where language could be produced with the desired politeness in order to influence or affect the audience’s behavior.
This study is also inspired by the work of Johnathan Culpeper [7, 6], he defined the impoliteness theory which mirrors the politeness theory but is defined in terms of impoliteness. The strategies used by him during his analysis of the Television (TV) Show “The Weakest Links” involves analyzing the response of a comment to try and predict whether the initial response was impolite or not. A few hypotheses in this work are based on the same strategy, even the data collection strategy is influenced by the same strategy.
There has been a lot of work done for developing politeness theories [4, 16, 6, 27]. These theories are based on various politeness and impoliteness strategies and are proven to work well in the real world. There has also been a lot of studies analyzing the linguistic aspects of social media. However, there has been no study which proves that these popular linguistics theories work on social media as well. This study tries to fill this research gap.

1.2 Research Question

The research question that the research is trying to answer in this study is:
"Does politeness theories which work well in real life, work well on social media forums also?"

1.3 Research Aim

The aim of this study is to study and use tools available to measure “Politeness Score” of a text in conjunction with politeness strategies to test whether popular politeness theories work in the social media world or not.

1.4 Hypotheses

Following hypothesis are tested in the study:

1. The politeness of comments on Stack Exchange is related to the reputation of the user who posts the comment.
2. The politeness of a comment is related to the number of upvotes or downvotes it gets.
3. While interacting with similar posts on Reddit socially powerful people generally get more polite replies to their comments when compared to replies on comments of people with less social power.
4. The politeness of a comment is based on the gender of the user posting it: in particular, females are hypothesized to write more polite comments.
5. The politeness of a reply made to a comment is based on the gender of the person who posted the comment: in particular, females get more polite replies to their comments.
6. Replies to comments made by users of opposite gender are more polite when compared with replies to comments posted by users of same gender.

1.5 Thesis Overview

This section describes the structure of the thesis. Thorough literature review is conducted in Chapter 2, Chapter 4 briefly describes all the datasets used in this study, Chapter 3 gives a thorough background of all the tools and techniques for this study, Chapter 5 describes each step taken for conducting this research in detail, Chapter 6 states and discusses the results, Chapter 7 concludes all the work done in this research.

Chapter 2

Literature Review

Politeness theory and its various aspects have been focused on by many researchers in the scientific community. Popular works on politeness theories include the works of Brown and Levinson [4], Johnathan Culpepper [7], Lakoff [16], Grice [14], and Watts [27]. These researchers developed several models and rules for politeness based on different theories.

In their work, Brown and Levinson theorized the concept of face and based on their theories they developed a universal model of linguistic politeness. Their work was one of the first to examine speakers from multiple languages for linguistic strategies used by them. They analyzed three languages English, Tzeltal, and Tamil. During their experiments, they observed that while the speakers of all the three languages make some particular types of requests or when these speakers criticize their hearers, they tend to use indirect speech or hedges more than any other form of speech. Before this research, Grice had theorized the maxims of manners which stated that speakers tend to avoid incomprehensibility, complexity, and vagueness in speech to achieve effective conversation in common social situations[14, 13]. The work of Brown and Levinson directly contradicted the maxims of Grice. Based on this contradiction, Brown and Levinson theorized that the use of these form of speeches (indirect speech and hedges) are used by speakers in order to make their speech more polite. On this notion, Brown and Levinson defined politeness and developed their theories of face around it [4, p. 55]. While developing their theories they also made use of positive and negative politeness strategies. Positive politeness strategies are a more direct form of expressing politeness, like using phrases like please and thanks, whereas negative politeness strategies are a more indirect form of expressing politeness, gestures like not interrupting someone when they are speaking. Erving Goffman developed the concept of face and facework [12], Goffman emphasizes on the significance of face as a discursive structure which is negotiated during the occurrence of social interaction. According to Goffman, during a conversation, face represents the social value that can be claimed by a person in front of others[12, p. 5]. He also suggests that face is a result of facework, which is defined as actions performed to make the face consistent[12, p. 12]. Brown

and Levinson matured this notion of face and theorized that this notion of face is the reason behind the politeness in any speech. Some researchers feel that the notion of face developed by Brown and Levinson is intellectual and theoretical but not pragmatic [10], however, their work is considered to be one of the most popular ones in this field and many researchers have developed their works by taking these theories as a base, for example, work done in [23], tries to find relationship between politeness and power status based on Brown and Levison's theories. Some works also modify and alter politeness theories proposed by Brown and Levinson so that these theories can adapt to real-life conversations [27, 11, 18].

In 1996, Johnathan Culpeper defined impoliteness theories which kind of mirrored politeness theories defined by Brown and Levinson [6]. These impoliteness theories are based on forms of speech such as sarcasm and mimicry. Similar to Brown and Levinson, Culpeper also defined his theories in terms of positive and negative impoliteness strategies. Later in another research, Culpeper analyzed some excerpts of scripts in the Television (TV) show "The Weakest Link" based on his different impoliteness theories like sarcasm and mimicry. He observed how these impoliteness theories are used in the TV show to create entertainment [7]. Majority of this work analyses how the comments made by the host are responded by the target, based on the responses and reactions of the target Culpeper analyses whether the original comment by the host should be considered impolite or not. For example, if the host of the TV show gets an angry response in between an interaction with the person in the crowd, Culpeper considered it to be logical that the initial comment to which the repose was made is impolite; similarly, if the host gets a good gesture (smiles, thank you, please, etc.) in response to his comment then it was considered not to be impolite by Culpeper. However, there were several factors which were not considered in the work, such as the factor that audience already has knowledge of the nature of the show, therefore, their responses and reactions might not be natural. Also, the presence of the camera might also influence the targets natural response. Another famous study is the work of Watts [27] who proposed a model of politeness theories using theories of social networks and practice.

Due to the emergence of the internet and social media, a large amount of digital data is available nowadays. This presents a lot of opportunity for research in the area of computation linguistics. Due to the availability of a large number of text corpora, there has been a lot of work done on analyses of language using computational approaches. Using the huge text corpora available on the internet, automation of the process of identification of politeness strategies is now possible theoretically. However, these automated tools still lag behind when it comes to interpreting the results, and hence these tools could right now only be used to recognize patterns in language using the large text corpora available and interpretations of these patterns and results still have to be conducted by the researchers. An era where machines can perform these kind of analysis is still far. According to Wiedemann, the main

advantages of these automated tools is that with the help of the huge corpora, they can address the problem of variability, consistency, and reliability of results of different studies. In his work, Wiedemann also states that as the scientific community is adopting these tools, the lines between qualitative and quantitative research are becoming more and more blur and the research community is heading into an era where it has researches which integrate both quantitative and qualitative aspects of research [28, p. 21].

Work performed by Danescu-Niculescu-Mizil is one of the first attempts to identify key linguistic features or markers for politeness[9]. Wikipedia edit requests and Stack Exchange comments and replies were used for data collection in this study. Amazon Mechanical Turk (AMT) has been used to manually annotate the data on their politeness scores. AMT provides human labor for various tasks like annotations, it is used in the scientific community to perform tasks which are not yet automated. For screening the annotators, a questionnaire was designed about linguistics. All annotators selected for this study had to go first pass the questionnaire. Each worker in AMT annotates 13 requests/comments and each request is annotated by 5 workers from 'very impolite' to 'polite'. Z-score normalization was done on each worker's score to standardize the score, as politeness is very subjective, and it is different for different people. Final politeness score for a request was then calculated using the average of the 5 normalized scores. Based on the annotated dataset and the politeness theories proposed by Brown and Levinson, Danescu-Niculescu-Mizil manually designed 20 lexical and syntactical features (markers) related to politeness. Using the annotated dataset, two Support Vector Machine (SVM) models were trained and compared based on their accuracies with reference to human accuracy in predicting politeness levels. The first SVM model used a Bag of Words (BOW) approach based on unigrams, the results of model state an accuracy of around 77% when trained and tested on the dataset of same domain (model trained on Wikipedia requests and tested on Wikipedia requests) and an accuracy of 69% when cross-domain testing is performed (model trained on Wikipedia requests and tested on Stack Exchange comments). The second model used the BOW approach based on unigrams and in addition to that uses the 20 lexical markers designed by Danescu-Niculescu-Mizil as features, by using these markers additional the accuracy of the model was increased to 81% when testing in the same domain and 72% when tested in cross-domain. The second SVM classifier is stated to achieve a near-human accuracy when tested in the same domain. The work also uses the politeness levels achieved by its model to determine the power of the user in the communities examined.

Malika and Mohit use the work of Danescu-Niculescu-Mizil et al. as a baseline and tries to improve the work by using Convolutional neural network (CNN) instead of SVM to increase the accuracy of measurement of politeness [1]. Instead of manually identifying the markers for politeness the work uses various techniques such as computer vision to perform qualitative and quantitative analysis of what is

being learned by the CNN w.r.t. politeness strategies to generate makers related to politeness. The generated markers and the analysis of learning of neural networks not only shows the politeness markers proposed by the work of Danescu-Niculescu-Mizil but also provides some interesting extensions to some of the existing markers. Also, some new markers were identified by CNN. These features when added in addition to the existing features to the SVM trained by Danescu-Niculescu-Mizil, increases the accuracy of SVM, suggesting they have good relevance to politeness.

Blog post in [2], summarizes a feature of IBM Watson which is in Beta stage. The feature analyses twitter conversations in customer support channels to predict tones. This work uses a methodology very similar to what Danescu-Niculescu-Mizil used. The tool trains a machine learning model based on a set of customer support conversations and their associated tones to predict tones for new customer support conversations in order to improve customer satisfaction. In this work, SVM is used as the machine learning model. The currently supported tones include frustration, sadness, satisfaction, excitement, politeness, impoliteness and sympathy. For training the machine learning model, several categories of features including n-gram features, lexical features from various dictionaries, punctuation, and existence of second-person references in the tone are leveraged. The system takes in each dialogue in a conversation as an input, based on the input a confidence score between 0 and 1 is generated by the system for each of the above mentioned 7 tones. The system only returns tones which have a confidence score higher than 0.5.

There has been a lot of work done on the development of politeness theories, some work has also been done on analyzing linguistic aspects over social media but there is a huge gap when it comes to testing these popular politeness theories over social media. This research tries to bridge this gap by hypothesizing various popular theories about politeness and testing these hypotheses over social media forums like Reddit, Stack Exchange and Wikipedia.

Chapter 3

Techniques and Tools used

This chapter gives the reader a background of tools and techniques which are used for this work in order to perform data collection, analysis, and hypothesis testing. Section 3.1 briefs reader about The Python Reddit API Wrapper (PRAW), Section 3.2 explains The Comprehensive R Archive Network (CRAN) politeness package, Section 3.3 explains the inner working of Least absolute shrinkage and selection operator (LASSO) regression, Section 3.4 walks reader through how Pearson Product-Moment Correlation Coefficient is used for calculating correlation between variables and performing hypothesis testing, and Section 3.5 explains how t-tests are used for hypothesis testing.

3.1 PRAW: The Python Reddit API Wrapper

PRAW is an easy to use python package that provides access to Reddit's API. All Reddit API rules are followed by PRAW and it is easy to install PRAW using the pip tool which is a standard package manager of python [22]. Several advantages of using PRAW include elimination of sleep calls in your code, getting structured data, etc. PRAW only requires Reddit client details and a Reddit useragent. The Reddit useragent follows the rules of Reddit API and PRAW handles these rules so that the end user does not have to worry about violating these rules[3]. PRAW has extensive tutorials and documentation, also many other researchers have been using PRAW in the scientific community for scrapping Reddit data [21, 19].

3.2 CRAN Politeness package

R is an open-source programming language and environment which provides tools for performing statistical and graphical techniques. CRAN is a web server and file transfer protocol (FTP) network that stores homogeneous and latest R distributions, R packages, R documents, R extensions, and R binaries. The CRAN politeness package is a tool which is developed by combining and extending previous work

on linguistic markers of politeness [5].

CRAN politeness package is a tool which is developed by combining and extending the available research on linguistic markers related to politeness [4, 9]. It is the first and most popular package in R to study politeness, it is also one of the first broad study of linguistic pragmatics in R. It provides tools for visualising, quantifying and evaluating politeness between different groups of text documents. This package is developed while keeping in mind to make it simple to recognize politeness in English language. It compares various other aspects of English language to quantified characteristics related to politeness in speech.

R package provides tools to extract politeness markers in English natural language. It also allows researchers to easily visualize and quantify politeness between groups of documents. This package combines and extends prior research on the linguistic markers of politeness.

The package provides functionality through which researchers can identify markers related to politeness in natural language. It also provides a graphical interpretation of these markers so that they can be visually compared to other covariates of interests. The package can also be used to train a supervised machine learning (ML) model to predict and detect politeness in new text documents. The politenessProjection function of the package gives the user functionality of developing a politeness classifier, it does that by creating a mapping between markers related to politeness in supplied text and other measures like politeness score or politeness levels. In other words, if some labeled data (labels of politeness over a set of text) is provided then a model can be trained using this labeled dataset to predict politeness labels of other texts. Markers related to politeness in the supplied text are used to train this model. The package uses 36 markers related to politeness, most of these markers are directly borrowed from recent literature and research on computational politeness. Table 3.1 lists down the politeness markers used by the package. The politenessProjection function works as a wrapper and uses supervised ML algorithms to train models for predictions. In default mode, Least Absolute Shrinkage and Selection Operator (LASSO) regression is used for making the predictions. Section 3.3 explains the working of lasso regression. The package also offers functions which provide insights such as texts which are most polite or least polite over a labeled dataset [29].

Feature Name	POS Tags	Description	Example
Hello	No	"hi", "hello", "hey"	"Hi, how are you today?"
Goodbye	No	"goodbye", "bye", "see you later"	"That's my best offer. Bye!"
Please Start	Yes	Please to start sentence	"Please let me know if that works"
Please	Both	Please mid sentence	"Let me know if that works, please"
Gratitude	Both	"thank you", "i appreciate", etc.	"Thanks for your interest"
Apologies	Both	"sorry", "oops", "excuse me", etc.	"I'm sorry for being so blunt"
Formal Title	No	"sir", "madam", "mister", etc.	"Sir, that is quite an offer."
Informal Title	No	"buddy", "chief", "boss", etc.	"Dude, that is quite an offer."
Swearing	No	Vulgarity of all sorts	"The dang price is too high"
Could You	No	Indirect request	"Could you lower the price?"
Can You	No	Direct request	"Can you lower the price?"
Bare Command	Yes	Unconjugated verb to start sentence	"Lower the price for me"
Let Me Know	No	"let me know"	"Let me know if that works"
Affirmation	Yes	Short appreciation to start sentence	"Cool, that will work out then"
Agreement	Yes	Explicit statement of agreement	"I mostly agree with that"
Acknowledgement	Yes	Explicit statement of understanding	"I understand your point"
Conjunction Start	Yes	Begin sentence with conjunction	"And if that works for you"
Reasoning	No	Explicit reference to reasons	"i want to explain my offer price"
Reassurance	No	Minimizing other's problems	"Don't worry, we're still on track"
Ask Agency	No	Request an action for self	"Let me step back for a minute"
Give Agency	No	Suggest an action for other	"I want to let you come out ahead"
Hedges	No	Indicators of uncertainty	"I might take the deal"
Actually	Both	Indicators of certainty	"This is definitely a good idea."
Positive	No	Positive emotion words	"that is a great deal"
Negative	No	Negative emotion words	"that is a bad deal"
Negation	No	Contradiction words	"This cannot be your best offer"
Questions	No	Question mark count	"Is this for real?"
WH Questions	Yes	Questions w/ WH words (how, why, etc)	"Why did you settle on that value?"
Yes/No Questions	Yes	Questions w/o WH words	"Is this for real?"
By The Way	No	"by the way"	"By the way, my old offer stands"
Adverbial Just	Yes	modifying a quantity with "just"	"It is just enough to be worth it"
Filler Pause	No	Filler words and verbal pauses	"That would be, um, fine"
For Me	No	"for me"	"It would be great for me"
For You	No	"for you"	"It would be great for you"
First Person Plural	No	First person plural pronouns	"it's a good deal for both of us"
First Person Single	Both	First person singular mid sentence	"It would benefit me, as well"
Second Person Single	Both	Second person mid sentence	"It would benefit you, as well"
First Person Start	Yes	First person singular to start sentence	"I would take that deal"
Second Person Start	Yes	Second person to start sentence	"You should take that deal"
Impersonal Pronoun	No	Non person referents	"That is a deal"

Table 3.1: Politeness markers used by CRAN politeness package

3.3 LASSO Regression

LASSO is a regression analysis method that performs both variable selection and regularization what it does is force the sum of the absolute value of the coefficients to be less than a fixed value which results in certain coefficients to be set to zero effectively removing those variables from the model making the model simpler and more interpretable the regularization parameter lambda (λ) governs the degree to which the coefficients are penalized.

LASSO regression is a linear regression that utilizes shrinkage. In shrinkage values in the dataset are shrunk to a point of interest like mean. As it sets the coefficient of less useful variables to zero, therefore, it promotes simple and sparse models. Lasso regression is preferred for models with high concentrations of multicollinearity, it also helps in automating the selection of model parts.

L1 regularization is conducted by LASSO regression. In L1 regularization a summation of the absolute value of coefficients is used as a penalty. A simpler and sparse models is generated by doing this

because large penalties of absolute values lead to coefficients values being close to zero.

The goal of LASSO regression is to minimize Equation 3.1

$$\sum_{i=1}^{n}(y_i - \sum_{j}^{p}(x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \tag{3.1}$$

In Equation 3.1 n is the number of observations in data, p is the number of variables in the model, β_j is the coefficient of j'th variable of the model, y_i is the the target value of i'th observation, x_{ij} is the value of j'th variable in the i'th observation, and λ is the tuning parameter. Equation 3.1 could also be realized as minimizing the sum of squares with a constraint of $\sum_j^p |\beta|_j \leq s$, where s is a slack variable. Lasso solutions are quadratic programming problems and after solving these problems many coefficients are shrunk to zero resulting in a simple and sparse model.

λ is the tuning parameter which controls the rate of shrinkage. In other works λ is a parameter which controls the penalty to the sum of squares. With increase in value of lambda more coefficients are shrunk to zero. λ controls the trade-off between bias and variance, if λ increases bias increase and on the other hand if λ decreases variance increases [26].

In Figure 3.1 it can be seen that when sum of squared residuals is minimized a very low bias is obtained but that results in high variance and generalisation over test data is not achieved, however with LASSO regression for a little bit compromise on the bias variance is reduced significantly.

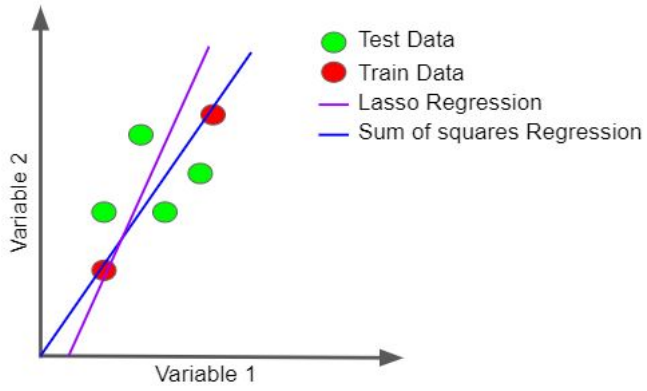

Figure 3.1: LASSO regression and sum of squares regression

3.4 Pearson Product-Moment Correlation Coefficient

Pearson Product-Moment Correlation Coefficient or Pearson's r correlation coefficient is used for finding the strength of the linear relationship between two variables. It measures the correlation between two variables by constructing a line of good fit through data points of the variables, and then the correlation is measured by calculating how far these data points are from this line of good fit.

The Pearson coefficient, r, can take a value between -1 to 1. Figure 3.2 shows how to interpret the Pearson's r coefficient. A positive value means that the variables are positively related, a negative value means that the variables are negatively related, while a value of 0 means the variables are not at all related. A value near -1 suggests a strong negative correlation while a value near +1 suggests a strong positive correlation. This means that more the magnitude of the correlation coefficient, r, the more the two variables are related, positively or negatively. Figures 3.5, 3.3 and 3.4 shows positively correlated data, negatively correlated data, and data which is not at all correlated respectively.

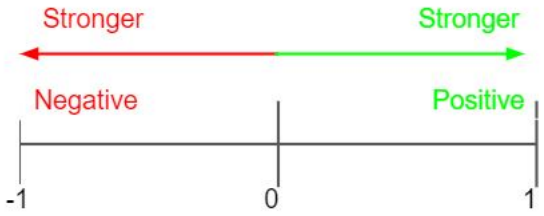

Figure 3.2: Interpretation of Pearson's r Coefficient

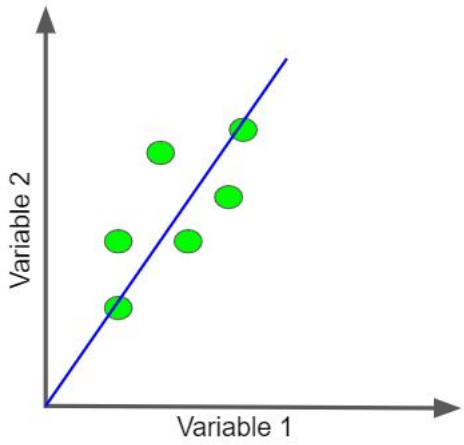

Figure 3.3: Positively correlated data

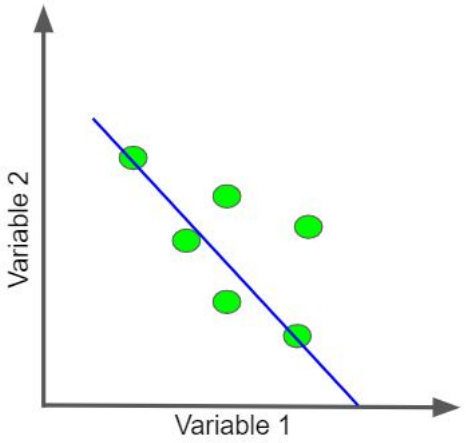

Figure 3.4: Negatively correlated data

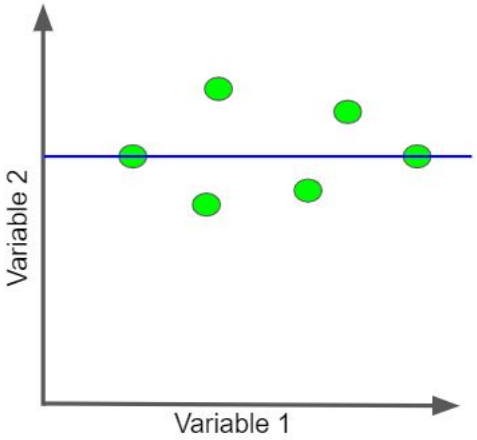

Figure 3.5: Data with no correlation

The Pearson's r correlation coefficient is calculated using the formula in Equation 3.2

$$r = \frac{N\sum_i^N (a_i b_i) - \sum_i^N a_i \sum_i^N b_i}{\sqrt{\left[N(\sum_i^N a^2) - (\sum_i^N a)^2\right]\left[N(\sum_i^N b^2) - (\sum_i^N b_i)^2\right]}} \tag{3.2}$$

In the Equation 3.2 a and b are the two variables, a_i denotes the i'th observation of a, b_i denotes the i'th observation of b, N is the number of pair of observtions, $\sum_i^N (a_i b_i)$ is the sum of the product of the paired observations, $\sum_i^N a_i$ is the sum of all observations in variable a, $\sum_i^N b_i$ is the sum of all observations in variable b, $\sum_i^N a^2$ denotes the squared sum of all observations in a, and $\sum_i^N b^2$ denotes the squared sum of all observations in b [20].

Pearson's r correlation could also be used for hypothesis testing. During hypothesis testing using Pearson's r correlation, null hypothesis is taken as $\sigma = 0$, where σ denotes the correlation between

the two variables. If the null hypothesis is rejected then it can be said that the two variables are statistically correlated, if rejection of the null hypothesis is failed then it cannot be said that the two variables are statistically correlated. Therefore, for performing hypothesis testing the hypothesis is set up as follows:

$$H_0; \sigma = 0$$

$$H_a; \sigma \neq 0$$

where H_0 denotes the null hypothesis, H_a denotes the alternative hypothesis, and σ denotes the correlation between the two variables.

After setting up our null and alternative hypothesis an alpha (α) value has sto be set, which denotes the significance level. Significance level denotes the probability of rejecting the null hypothesis when it is true. Usually, researchers use an α value of 0.05, other common values of α are 0.1 and 0.01. For testing our hypothesis degrees of freedom is also needed. Degrees of freedom is nothing but the number of observations minus the number of variables, for Pearson's correlation number of variables are 2, therefore, degrees of freedom is $n-2$, where n is the number of observations in the dataset. Once the degrees of freedom is decided upon in conjunction with an α value then a critical value can be looked for in the r tables. Critical value is the deciding factor for our decision rule. If the correlation score, r, is more than the critical value then the null hypothesis is rejected, H_0, meaning that correlation in the two variable is statistically significant, if the r score is less than the critical value then rejection of the null hypothesis fails.

3.5 T-test and F-test

Purpose of this section is to explain t-tests how they work and when to use them. A t-test checks the averages or means of two groups are reliably different. Instead of using t-tests, we may just look at the means, looking at the means may tell us if there's any difference at all but that doesn't tell us if the difference is reliable. For example, if we flip two coins 100 times and we get heads 52 times on one coin and get heads 49 times on the other coin that does not explain that we reliably get more heads on one coin or is one coin somehow more likely to get heads in the future. By looking at the means we can not say if there's a significant difference between them. This is where the difference between descriptive and inferential statistics come into place.

A descriptive statistic only describes the sample we have it doesn't tell us if our results are likely to happen again in contrast. A t-test is what we call an inferential statistic, inferential statistics don't just describe our sample they tell us what we can expect in new samples that we don't even have. Inferential statistics allow us to generalize our findings to a whole population beyond the sample that

we're testing that can be very powerful. In our example of coin flips the numbers 52 and 49 are different but does one coin actually is more likely to get heads or was this just chance? Would a similar result happen again with a new sample? To answer these questions we need inferential and not just descriptive stats. The t-test will tell us how likely this difference is to be reliable or whether it's just due to chance.

T-test measures the difference between the groups and compares it to the difference within the groups. The t-value is just a ratio of the variance between groups and variance within groups. A t-value of three into two groups are about three times as different from each other as they are within each other. That also means that if groups have wider more scattered scores it will be harder to detect a real difference between the groups and if they had narrow tightly clustered scores then the difference is easier to detect[25].

Like we had a significance level, α, associated with each Pearson's r score in Section 3.4, here also each t value has a corresponding p-value. The p-value is the probability that the pattern produced by our data could be produced by random data in other words it tells us whether the difference between our groups is real or if it's just a fluke so a p-value of 0.05 means there's only a 5% chance we would get these results with random data a p-value of 0.01 means there's only a 1% chance we would get these results with random data while point 1 means there's a 10%.

The exact p-value associated with the t-value depends on how many people are in your sample bigger samples make it easier to find statistically significant differences for example with two groups of five a t-value of 2 has a p-value of 0.05 when you increase the sample size to two groups of 10 that same t value of 2 now has a p-value of 0.03 bigger samples are helpful but the benefit diminishes as the sample size increases. If the sample is too small, we might not have the statistical power to detect the present differences. Similar to Pearson's correlation in Section 3.4, the sample size of the dataset is represented through a number called degrees of freedom. For t-tests, the degree of freedom is the sample size minus 1.

There are three main types of t-test the independent samples, paired samples, and the one-sample test. The most common type is the independent samples t-test, this is when you have two different groups you want to compare. Another type of t-test is the paired-sampled t-test this is we have one group that is measured at two different times, in a paired samples t-test each score is paired with another score usually because the measurements come from the same subject this is different from an independent sample t-test where scores between groups are not related. The last type of t-test is the one-sample t-test this is when we only have one group and we want to compare it to a hypothetical value or a known population mean. In this study, we only focus on the independent sample t-test.

The independent sample t-test is again of two types, one-tailed t-test and two-tailed t-test. Two-tailed

t-test tests the statistical significance in both the direction, it allocates half of the decided p-value at one end and half of the decided p-value at the other end. For a p-value of 0.05, it allocates 0.025 at one end and 0.025 at the other end. This means that regardless of the hypothesis it checks whether the means are significantly greater or significantly smaller from one another. The alternative hypothesis in a two-tailed t-test is that the means are unequal. Whereas, the one sampled t-test checks statistical significance in only one direction. It allocates all of the decided p-value at one end only. Since it checks statistical significance in one direction, it can only check whether the means differ from one another in one direction, i.e, it can either check that one mean is greater than other mean or it can check that one mean is less than the other mean. The alternative hypothesis in one-tailed t-test either states that one mean is greater than the other mean or it states that one mean is less than the other mean. In this study, we only focus on the two-tailed t-test.

The two sampled t-test are again of two types, two-tailed t-test assuming equal variances and two-tailed t-test assuming unequal variances. If the variances of two variables are proven to be statistically equal then we employ t-test assuming equal variance otherwise we employ t-test assuming unequal variances. Equation 3.3 gives the formula of t value assuming equal variance.

$$t = \frac{\bar{a} - \bar{b}}{s^2 \sqrt{\frac{1}{n1} + \frac{1}{n2}}} \tag{3.3}$$

$$\text{where, } s^2 = \frac{\sum_{i=1}^{n1}(a_i - \bar{a}) + \sum_{i=1}^{n2}(b_i - \bar{b})}{n1 + n2 - 2} \tag{3.4}$$

In the Equation 3.3, $\bar{a}$ and $\bar{b}$ are the means of the variables, $n1$ and $n2$ are the number of observations under variables a and b respectively, s^2 is the pooled sample variance.

Equation 3.5 gives the formula of t value assuming unequal variance.

$$t1 = \frac{\bar{a} - \bar{b}}{\sqrt{\frac{s1^2}{n1} + \frac{s2^2}{n2}}} \tag{3.5}$$

$$\text{where, } s1^2 = \frac{\sum_{i=1}^{n1}(a_i - \bar{a})}{n1 - 1} \tag{3.6}$$

$$\text{and, } s2^2 = \frac{\sum_{i=1}^{n2}(b_i - \bar{b})}{n2 - 1} \tag{3.7}$$

In the Equation 3.5, $\bar{a}$ and $\bar{b}$ are the means of the variables, $n1$ and $n2$ are the number of observations under variables a and b respectively, $s1^2$ and $s2^2$ are the sample variances of variables a and b respectively. Degrees of freedom is also calculated differently in case of t-test assuming unequal variances. Equation 3.8 gives the formula for calculating the degrees of freedom in case of t-test assuming unequal

variances.

$$df = \frac{\left[\frac{s1^2}{n1} + \frac{s2^2}{n2}\right]^2}{\frac{(\frac{s1^2}{n1})^2}{n1-1} + \frac{(\frac{s2^2}{n2})^2}{n2-1}} \tag{3.8}$$

For checking whether the variances of the two variables are equal statistically we employ F-test. F-test can also be one-tailed and two-tailed. For the purpose of this study, we only focus on two-tailed t-test. The hypothesis for two-tailed F-test is given as follows-:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_a : \sigma_1^2 \neq \sigma_2^2$$

where σ_1^2 and σ_2^2 are the population variances of the two variables. The test statistic for F-test is the ratio between the sample variances of the two variables. It is defined in Equation 3.9.

$$\text{F-score} = \frac{v_1^2}{v_2^2} \tag{3.9}$$

In 3.9 $v1^2$ and v_2^2 are the sample variances of the two variables. The probability of two variables having unequal variances depends on the F-score, more the F-score deviates from 1 more is the probability of the variables having unequal variance. Similar to the t-test F-test also has a p-value or significance level associated with it. The null hypothesis that the two variances are equal is rejected if F-score of the sample variances is greater than the F-score given the significance level and degrees of freedom. The degrees of freedom are defined as the number of observations minus one for F-test. Equation 3.10 defines when a null hypothesis is rejected in the F-test.

$$\text{F-score}_{v_1^2, v_2^2} > \text{F-score}_{p-value, N1-1, N2-1} \tag{3.10}$$

In Equation 3.10 v_1^2 and v_2^2 are sample variances for the two variables, p-value is the significance level for the test, $N1$ and $N2$ are the number of observations for the two variables [24].

Chapter 4

Dataset Description

High quality labelled data is required for performing computational studies of any facet of linguistics. This section discusses the datasets that are used for performing the analyses in our work. The study makes use of three datasets for testing all our hypotheses.

- Politeness annotated Wikipedia edit requests and comments on Stack Exchange. [9, 8]
- Reddit Comments, Replies and Karma dataset. [15]
- Reddit rateme subreddit Gender statistics dataset. [17]

Manually annotated data of edit requests on Wikipedia and comments on Stack Exchange are borrowed directly from [9]. We used and manipulated this publicly available dataset for training our politeness model. Section 4.1 explains more about this dataset and how it has been used in our analyses. For testing the hypothesis in this work which are based on social power, data collection over a popular web discussion forum Reddit is carried out. Section 4.2 elaborates more on the data collection strategy and the structure of the collected data. For our hypothesis which are based on gender of a person, the work uses the dataset available on Kaggle as Reddit provides gender for some subreddits and some users only. The dataset on Kaggle has comments and replies on several posts on rateme subreddit with the information on user's gender. This dataset was modified so that it can be used to test our hypothesis based on gender. The dataset itself and the modifications performed for using this data to test our hypothesis are explained in Section 4.3

4.1 Politeness annotated Wikipedia edit requests and comments on Stack Exchange

To evaluate and uncover various politeness theories, a new dataset of request annotated for politeness was proposed in the work of [9]. This dataset is used as a standard dataset for many other studies. The

dataset consists of data collected from two different sources Wikipedia and Stack Exchange. Requests plays a very important role in both of these online platforms.

Wikipedia provides user talk pages where publishers can communicate with each other to coordinate the development and maintenance of the cooperative encyclopedia. The requests on these talk pages are publicly available but generally these requests are targeted on the owner of the talk page. Users on Stack Exchange often ask for additional knowledge or suggests edit on the original posts via the medium of comments. Here also these comments are generally targeted on the owner of the post.

Both of these platforms are not only wealthy in requests engaging various users, these requests also result into consequential discussions minimising social repartee. Amazon Mechanical Turk (AMT) was used to annotate more than 1000 utterances of these requests, resulting in the largest dataset of politeness annotated data. Each request is annotated by 5 AMT workers and each worker annotated 13 requests. As politeness is very subjective so it is natural that each worker has his own scale for measuring politeness, hence, score of each AMT worker was normalised using z-score normalisation. After normalising each worker's score, final politeness score of each request is determined by taking average of each worker's score.

In this work, the comments on Stack Exchange are used to perform some initial analysis based on the reputation, up-votes, down-votes and the assigned politeness score of each comment. There were some manipulations to be done on the original CSV files to perform these analysis. The original Stack Exchange data was contained into two different CSV's. One CSV file contained information such as reputation of the user, up-votes and down-votes for each comment, whereas, information related to politeness score of the comments was contained in another CSV files, however a field named 'ID' connected both of these files. In order to bring all the information in the same CSV and to perform the required analysis a VLOOKUP querry was performed. Figures 4.1 and 4.2 show snapshots of original CSV files containing StackExchange comments. Figure 4.3 shows a snapshot of the final CSV.

Section 5.1 describes the initial analysis done on this dataset

Community	Id	Request	Timestam	UserId	Reputation	Upvotes	Downvote
092011 Androi	36	@JadeMason: In the Goo	2010-09-1	45	858	8	6
092011 Androi	50	I'm running 2.2 unrooted	2010-09-1	36	1021	57	2
092011 Androi	59	I have been thinking abou	2010-09-1	100	138	16	0
092011 Androi	63	What statistics are you re	2010-09-1	102	298	10	0
092011 Androi	78	I don't see "Set alarm..." i	2010-09-1	116	742	44	1
092011 Androi	83	I have a Milestone and I g	2010-09-1	105	546	18	2
092011 Androi	250	I'm not sure what you me	2010-09-1	47	494	41	1
092011 Androi	259	There was a known "virus	2010-09-1	49	497	74	0
092011 Androi	325	How are you trying to tet	2010-07-1	20	688	373	0

Figure 4.1: Snapshot of CSV file containing Reputation, Up-votes and Down-votes of StackExchange comments

Community	Id	Request	Score1	Score2	Score3	Score4	Score5	TurkId1	TurkId2	TurkId3	TurkId4	TurkId5	Normalized Score
092011 Stack	3210318	Can you explain more in	12	16	15	16	13	A33SMNN	A2OXXHG	A28TXBSZ	A3EJ5TT2	A3OY0OL2	0.217326
092011 Stack	1314703	Will expressions always b	13	13	18	14	12	A1UIH2IM	A23FB7HE	A17G2LOY	A2BKPNKL	A3L459MS	0.063302
092011 Stack	4740262	how are you resolving fur	13	15	17	13	17	A1BS64O3	A2AE4MZ'	ARJ44YPG	A14TK8NC	A17G2LOY	0.128902
092011 Stack	6972808	What is the definition of `	13	13	13	13	19	A3VJDU2V	A1ZHP80C	A28TXBSZ	A3MMLCE	A2BKPNKL	0.240188
092011 Stack	5269106	Is `A` a global variable? V	17	18	16	16	13	A3OW54N	A2RDZ580	A872FSFU	A1HBDQ0	A34M93N	0.508284
092011 Stack	7118357	This is a very confusing q	16	14	1	14	9	A2WKPCZI	A1BS64O3	A2USG66F	AYG3MF0	A1G05O3I	-0.39362
092011 Stack	5245510	Why not using `isnan()` fr	7	20	15	7	13	AL97SCCN	A3E157ZN	A2OTEOJC	A14TK8NC	A2GSW5R	-0.6897
092011 Stack	7628452	@Foon I just know that t	13	9	15	17	18	A416CQ0F	A3OW54N	A28TXBSZ	A3MMLCE	A1KC138N	0.178819
092011 Stack	7907868	do you want to send the	17	13	17	13	13	A2AE4MZ'	AYG3MF0	ARYGQ46I	A3K2Y59J	A1IMICF9	0.203021

Figure 4.2: Snapshot of CSV file containing information about annotated politeness score of StackExchange comments

Community	Id	Request	Score1	Score2	Score3	Score4	Score5	TurkId1	TurkId2	TurkId3	TurkId4	TurkId5	Normalized Score	Reputation	Upvotes	Downvotes
092011 Stack	3210318	Can you explain r	12	16	15	16	13	A33SMNN	A2OXXHG	A28TXBSZ	A3EJ5TT2	A3OY0OL2	0.217325858	1923	237	3
092011 Stack	1314703	Will expressions a	13	13	18	14	12	A1UIH2IM	A23FB7HE	A17G2LOY	A2BKPNKL	A3L459MS	0.063301586	38616	2164	328
092011 Stack	4740262	how are you resc	13	15	17	13	17	A1BS64O3	A2AE4MZ'	ARJ44YPG	A14TK8NC	A3L459MS	0.128901615	139	24	1
092011 Stack	6972808	What is the defin	13	13	13	13	19	A3VJDU2V	A1ZHP80C	A28TXBSZ	A3MMLCE	A3L459MS	0.240187562	12303	212	182
092011 Stack	5269106	Is `A` a global var	17	18	16	16	13	A3OW54N	A2RDZ580	A872FSFU	A1HBDQ0	A3L459MS	0.508284429	9196	1052	12
092011 Stack	7118357	This is a very con	16	14	1	14	9	A2WKPCZI	A1BS64O3	A2USG66F	AYG3MF0	A3L459MS	-0.393623139	1781	129	5
092011 Stack	5245510	Why not using `is	7	20	15	7	13	AL97SCCN	A3E157ZN	A2OTEOJC	A14TK8NC	A3L459MS	-0.689700981	558	149	14
092011 Stack	7628452	@Foon I just kno	13	9	15	17	18	A416CQ0F	A3OW54N	A28TXBSZ	A3MMLCE	A3L459MS	0.17881902	11	0	0
092011 Stack	7907868	do you want to se	17	13	17	13	13	A2AE4MZ'	AYG3MF0	ARYGQ46I	A3K2Y59J	A3L459MS	0.203020969	3522	154	48

Figure 4.3: Snapshot of the final CSV file containing all information of StackExchange comments such as Up-votes, Down-votes, Reputation, politeness scores etc.

4.2 Reddit Comments, Replies and Karma dataset

Reddit is a collection of online discussion forums where individuals can share information (news, content, knowledge etc.) and comment on posts from other individuals. For each user Reddit uses Karma as a scoreboard, user karma depends on the number of up-votes or down-votes he has on his posts and comments. Up-votes on posts or comments increase Karma, whereas, down-votes on posts or comments decrease user Karma.

Concept of Karma is useful for testing our hypothesis as it directly relates to the social power of a user on Reddit. To perform our analysis and test our hypothesis we collected data from 4 different subreddit forums ('Linguistics', 'Python', 'Words', 'WorldNews'). Figure 4.4 shows how Reddit is structured.

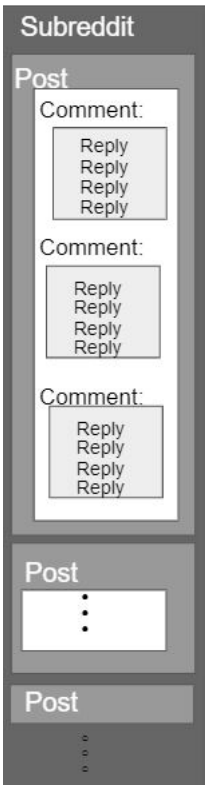

Figure 4.4: Reddit Structure

Reddit has a number of subreddits, each subreddit is related to a topic of discussion. In our analysis, this not only helps in limiting the point of the discussion to a particular topic but also gives the same reference point for measuring politeness score, as posts or comments in one subreddit can be more polite or impolite than posts or comments in another subreddit. Since the topic of discussion is same, therefore we minimise the chance of user's communicating on different politeness reference.
Each subreddit provides a space for users to post information related to a certain topic. Each of these posts can receive many comments from other users who share and discuss their thoughts on the posts via comments. Each comment can also have replies to it.
To test our hypothesis we need to collect comments, information about the user who made this comment and replies to these comments. In his work, Culpeper analyses some excerpts of scripts in the "The Weakest Link" based on different impoliteness theories like sarcasm and mimicry and observe how these theories are used in the TV show to create entertainment [7]. Majority of the work analyses how the comments made by host are responded by the target, based on these responses and reactions author makes a judgement whether the original comment by the host should be considered impolite

or not. Our data collection strategy was inspired by the work of Culpeper. Figure 4.5 shows the data collection strategy.

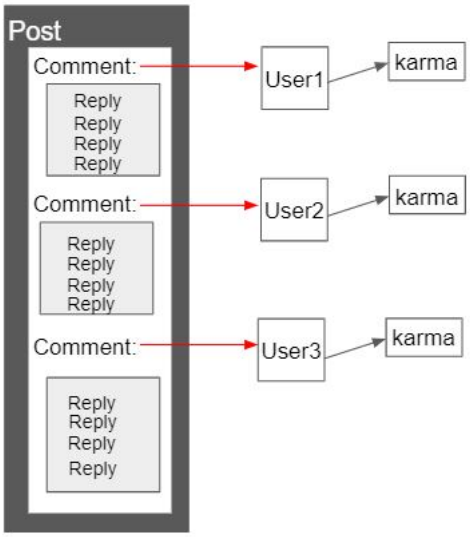

Figure 4.5: Data Collection Strategy

We collected comments and replies on several posts in a subreddit. For each comment we collect the relevant information of the user (karma, gender, etc.). We preformed the same data collection strategy for 4 subreddit, this was done in order to not restricted our hypothesis to one subreddit. Figure 4.6 shows the snapshot of data after collection. Section 5.4 explains about the analysis performed using this Reddit data.

CommentNumber	Comment	karma	Replies
0	Well the black	10993	Papiamento is a Portuguese creole spoken in some Carribean islands (mainly CuraÃ§ao and Aruba)
0	Well the black	10993	In
0	Well the black	10993	Costa Rica has Mekatelyu, which is very similar to Jamaican Creole.
1	In Brazil, we	10699	Can confirm. Am from Rio and pronounce *histÃ³ria* as â€œshtÃ³riaâ€
1	In Brazil, we	10699	I was in South Brazil on a short visit around thirty years ago, and people there thought I was a descendant of German people living in sc
2	On the island o	1637	CuraÃ§ao and Bonaire have Papiament*u*, and on Aruba it's Papiament*o*.
3	In Brazil	10709	While it may be the case that "there's no huge, culturally recognized dialect that just sounds 'black'", there absolutely is such thing as Af
4	Look into	12261	From my phonetics class in Spanish, Dominican, Cuban and Puerto Rican Spanish largely sound similar to Spanish from Andalusia, many
5	I once saw an	12728	> I once
5	I once saw an	12728	Would you happen to know the original Spanish title?
7	https://www.	78	Wow, this author actually writes some pretty good pop ling stuff! Any time someone goes right to Penny Eckert to talk about the CVS, yc
7	https://www.	78	This is the one I read! I meant it as a basis for the type of analysis -- not to say that it's the same dialect as indie voice. Sorry to be unclea
8	Iâ€™d love to s	307	Listen to I Miss You by blink-182, or any song by Neck Deep. The singers sing in an exaggerated SoCal, whiny, high pitched voice.
9	[There's this	61971	[deleted]
10	fun note I call t	8817	Shawn Mendes is a queer woman?

Figure 4.6: Snapshot of data after collection from Reddit

4.3 Reddit rateme subreddit gender statistics dataset

For testing our hypothesis which correlates the gender of the user to the politeness of his comments or the replies a user gets on his comments we needed the user gender information. Reddit provides user gender information only on specific subreddits and that too only for users who want to share their gender on these subreddit forums. So normal data scrapping techniques could not have been used to collect the gender of a person. One strategy could have been to collect the comments of users whose gender are known but these comments could have been posted in various subreddits and the reference point of these comments would not have been the same. It is also not necessary that replies to these comments have gender associated with them.

Keeping the above constraints in mind it was more logical to use an available dataset which contained comments, replies to these comments and the gender of the user who made the comment or reply. [17] is a dataset on Kaggle which contains all posts, comments, and replies to these comments from starting date of the subreddit rateme till January 2018. The dataset is a perfect fit for our analysis as it contained all relevant information which is needed to test our hypothesis about the majority of the users (gender, age, etc.) who have posted, commented or replied on this subreddit. The dataset is huge and has many missing values, it was modified accordingly to make it suitable for our analysis. The initial dataset has 294957 rows and 34 columns. After preprocessing the data according to the requirements of hypothesis testing we had 9256 rows and 7 relevant columns.
The data preprocessing strategy was to use comments who have a parent comment, i.e., to collect replies instead of comments. This was done in order to facilitate our hypothesis which are based on replies to comments. Then only those replies were filtered which had the gender of the user who made the reply and the gender of the user to which the reply was made to. Figure 4.7 shows the final dataset after performing the preprocessing and the analysis. Comment_body has text of the reply, comment_id had id of the reply, parent_comment_id has id of the comment to which the reply was made to, comment_author_gender has gender of the user who made the reply, comment_parent_gender has the gender of the user to which the comment was made to, column same_opp has the value 'same' if the gender of the user who made the reply and the gender of the user to which the comment was made to are same, if the gender of the user who made the reply and the gender of the person to which the reply was made to is opposite then the column has the value 'opp'. Finally, the predictions column has the predicted politeness score of the reply.
Section 5.5 elaborates on the analyses performed on this preprocessed dataset.

	comment_body	comment_id	comment_parent_id	comment_author_gender	comment_parent_gender	same_opp	predictions
1	Thanks. :) Pictur	c14xk1h	c14xfh9	male	male	same	0.736252
2	Actually. I just	c14xl5w	c14xk1h	male	male	same	-0.01185
3	Haha thanks. Ye	c14ychc	c14xlqa	male	male	same	0.924849
4	Let's test that	c1bol7n	c1bognn	male	female	opp	0.295678
5	I like it. and I lik	c1bp0uo	c1bol7n	female	male	opp	0.317151
6	I was definitely	c1bpnm3	c1bol7n	female	male	opp	0.278984
7	Yeah. I'm	c1ms93w	c1m8lgs	female	male	opp	0.394734
8	heh thanks	c1m7jga	c1m7i5i	female	male	opp	0.847689
9	oh and I do hav	c1mc1wa	c1mc0qp	female	female	same	0.161622
10	Thank you ! I ap	c21kcw9	c21hp99	female	male	opp	1.003218

Figure 4.7: Snapshot of preprocessed rateme subreddit gender statistics dataset

Chapter 5

Methodology

In this section we discuss the methodology adapted to perform the analyses for testing our hypotheses. The analysis can be broken down into five different sections.

- Initial Analysis on preprocessed Wikipedia edit requests and Stack Exchange comments dataset.
- Using python PRAW package for collecting Reddit data.
- Training model for predicting politeness scores using CRAN politeness package.
- Analysis on Reddit Comments, Replies and Karma Data and their predicted politeness scores.
- Analysis on preprocessed rateme subreddit gender statistics dataset and their predicted politeness scores.

Figure 5.1 shows the methodology of the work performed.

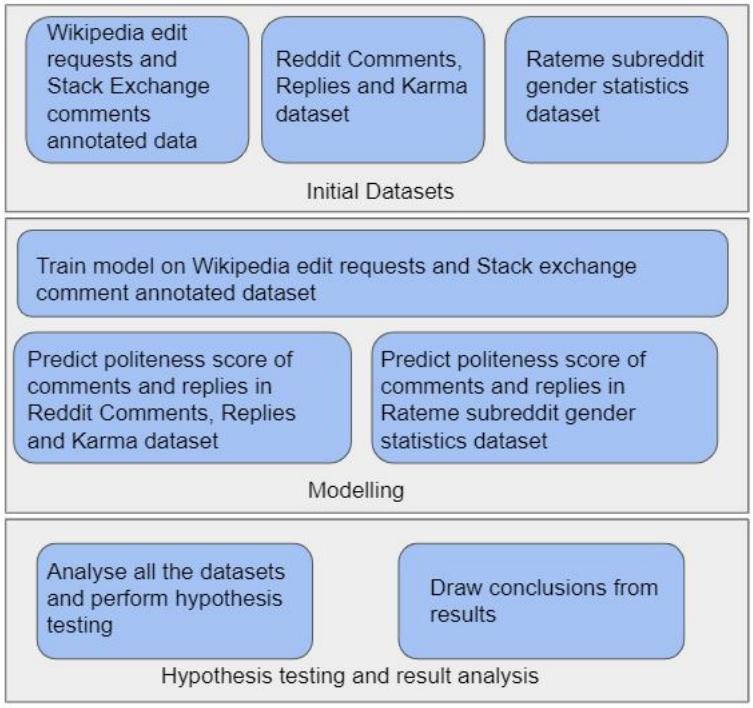

Figure 5.1: Methodology

Section 5.1 elaborates on how Stack Overflow data was analysed as per our hypotheses to obtain some interesting results. Section 5.2 expands on how PRAW was used to collect data from Reddit as per the requirement of our hypothesis. Section 5.3 briefs readers about how CRAN politeness package is used in this work for predicting politeness score of the replies and comments collected from Reddit. Section 5.4 expatiates on what analyses were used to analyse the data collected from Reddit. 5.5 dwells into analysis of rateme subreddit gender statistics dataset.

5.1 Initial Analysis on preprocessed Wikipedia edit requests and Stack Exchange comments dataset.

In Section 4.1 preprocessing of Wikipedia edit requests and Stack Exchange comments dataset has been explained in detail. In this section we explain how this preprocessed dataset has been analysed to perform hypothesis testing. We tested two hypotheses using this dataset. 4.3 shows the preprocessed dataset on which hypothesis testing is conducted. Figure 5.2 shows the histogram of politeness scores on the comments. From Figure 5.2 we can see that the distribution of the dataset is normal, this is because of the z-score normalization which is applied to the annotators scores as explained in Section 4.1. Also, it is very intuitive that this histogram should form a normal distribution as most of the comments would be moderate and some very impolite and some very polite. This intuition can be

confirmed by seeing Figure 5.2 which peaks at a politeness score of 0.

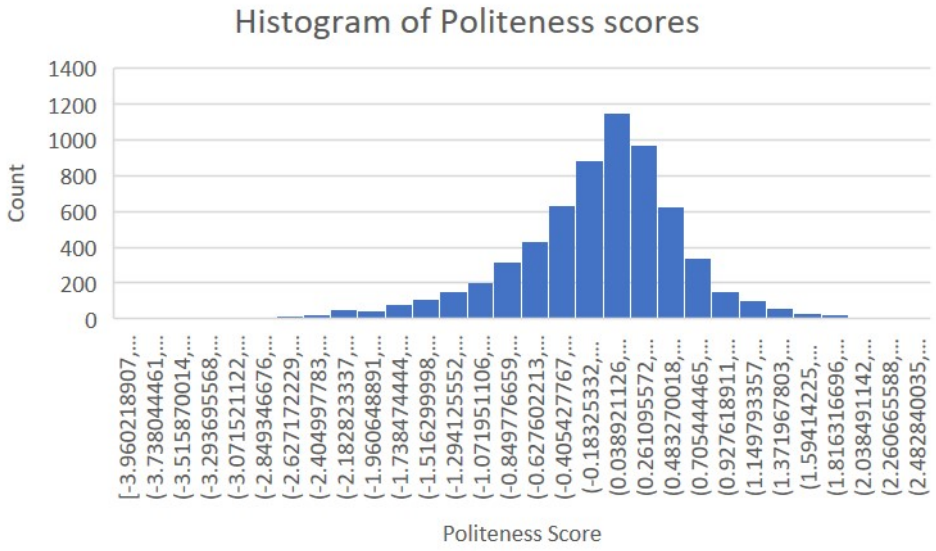

Figure 5.2: Histogram of manually annotated politeness scores of comments

Table 5.1 shows the descriptive statistics of the politeness scores.

Statistic	Value
mean	0.000381493
median	0.090210881
mode	-0.112877271
variance	0.50172326
standard deviation	0.708324262

Table 5.1: Descriptive statistics for politeness scores of comments

From the Table 5.1 we can see that mean, median and mode are all close to 0 which makes sense as the histogram curve also peaks at 0. Also, variance and standard deviation values are reasonable. These results are majorly because of the z score normalization.

Section 5.1.1 explains our analysis on the hypothesis that the politeness of a comment is related to the reputation of the user who posted the comment. Section 5.1.2 briefs the reader about our analysis on the hypothesis that the politeness of a comment is related to upvotes or downvotes it gets.

5.1.1 Analysis on Hypothesis 1

The hypothesis we are trying to test here states that the politeness of a comment is related to the reputation of the user who posted the comment. Figure 5.3 shows the hist of the reputation of users on Stack Exchange. From Figure 5.3 we can see that most of the users have reputation scores between 1-1000.

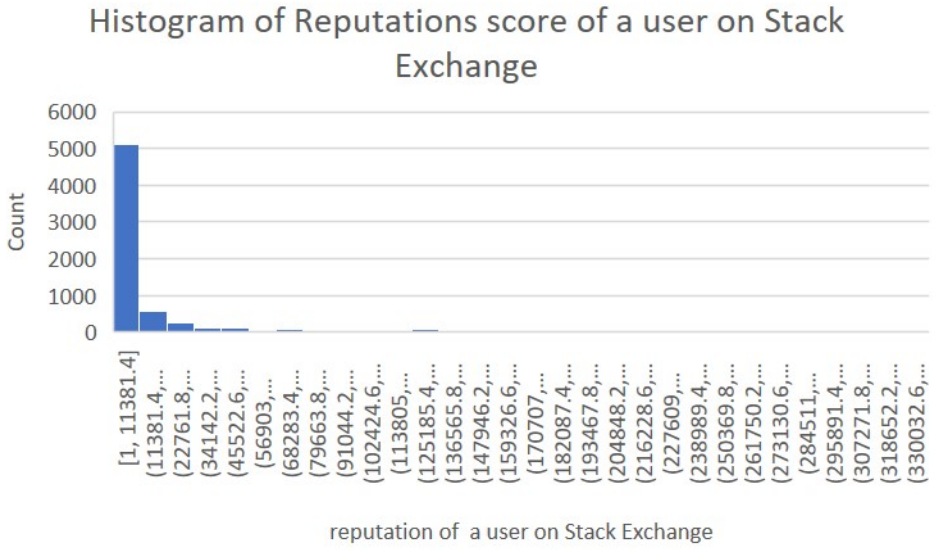

Figure 5.3: Histogram of reputation of the users

Table 5.2 shows descriptive statistics for the reputation of users who made the comment. We can see in Table 5.2 that mean, median and mode differ quite significantly. Also, variance and standard deviation is quite high. This is mainly because of the data has a lot outliers. We also need to consider that Pearson's r correlation does not work well with data which has a lot of outliers. We need to drill down our data in order to perform outlier removal.

Statistic	Value
mean	11317.67153
median	2328.5
mode	322
variance	705502280.2
standard deviation	26561.29289

Table 5.2: Descriptive statistics for reputation of users who posted comments

If we drill down the histogram to only those users who have reputation of 1000 or less, we get the

histogram shown in Figure 5.4. We can see from the histogram that after considering only comments whose author has reputation of less than 1000 we get a more uniform curve, there is also a hind of normality in the distribution.

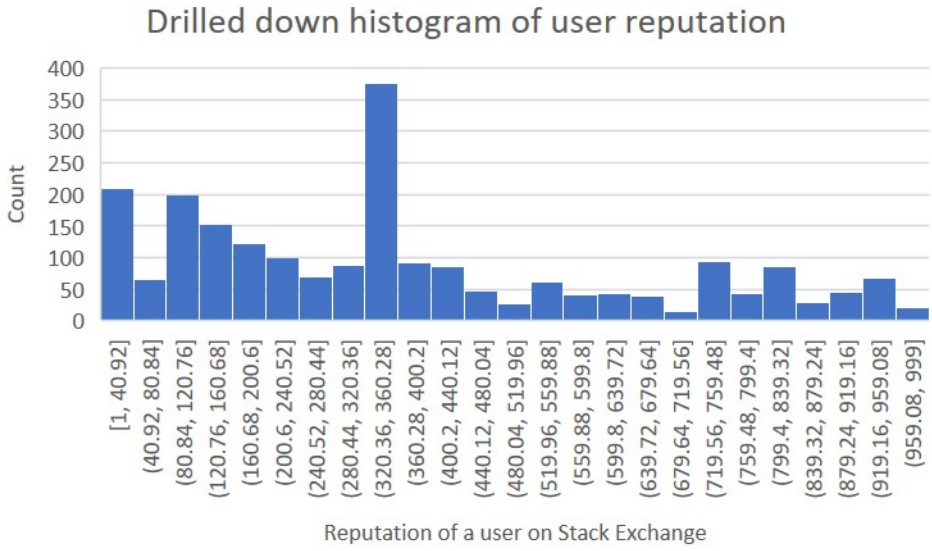

Figure 5.4: Histogram of reputation of the users after removing outliers

Table 5.3 shows descriptive statistics of reputation of users after outlier removal. We can see from Table 5.3 that mean, median and mode are quite close to each other and we have bought down the variance and standard deviation significantly. This would be a good dataset for performing Pearson's r correlation.

Statistic	Value
mean	364.4254937
median	322
mode	322
variance	72205.14891
standard deviation	268.7101578

Table 5.3: Descriptive statistics for reputation of users who posted comments after removing outliers

Figure 5.5 shows the scatter plot between politeness scores and reputation of comments (without removing outliers), from the Figure we cannot see any apparent correlation between these two variables, even the trend line in the figure does not show any significant relationship. However, there is a slight downward trend in both the scatter plots. This downward trend indicates a slight negative

correlation.

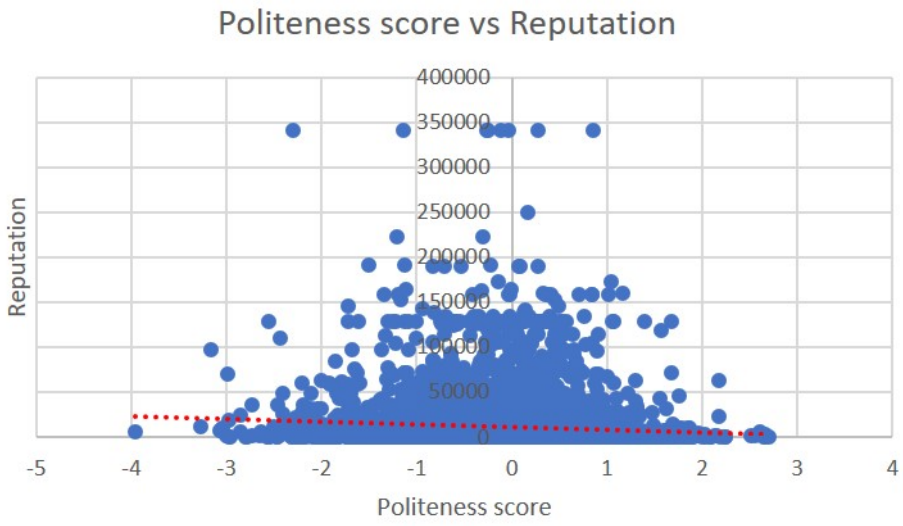

Figure 5.5: Scatter plot between politeness score of the comment and reputation of the author of the comment

We have used Pearson's r correlation on both the datasets, before and after outlier detection. Person's r correlation is explained in Section 3.4. A Pearson's r correlation was calculated between the politeness score of the comment and the reputation of the user who has posted the comment. We set our significance level, α, as 0.05. The results of the Pearson's r correlation and our analysis are explained in Section 6.1.1. We have rejected the hypothesis based on the results from our analysis.

5.1.2 Analysis on Hypothesis 2

The hypothesis we are trying to test here states that the politeness of a comment is related to the upvotes or downvotes it gets. Figure 5.6 shows the histogram of the upvotes of comments from Stack Exchange. From Figure 5.6 we can see that most of the comments get less than 1500 upvotes. Figure 5.7 shows the histogram of the downvotes of comments from Stack Exchange. From Figure 5.7 we can see that most of the comments get less than 303 downvotes.

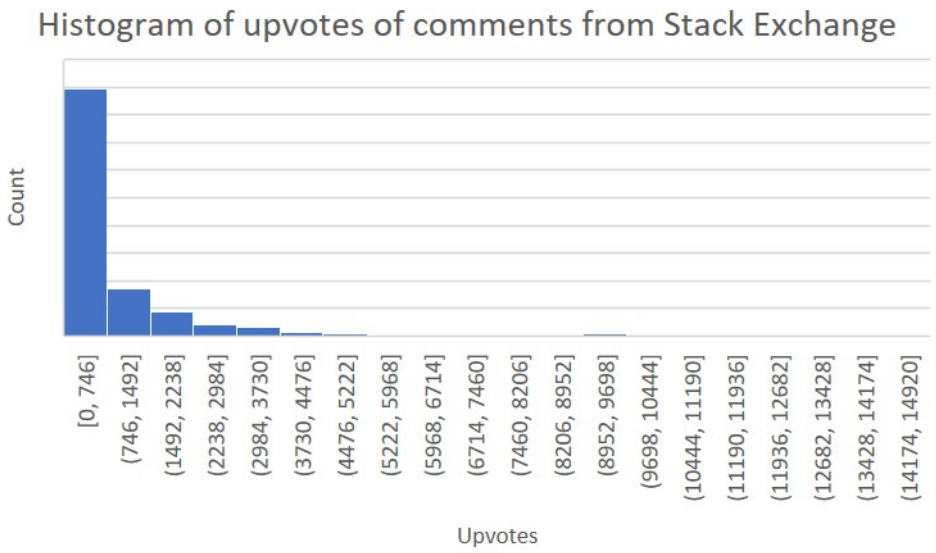

Figure 5.6: Histogram of number of upvotes on all comments

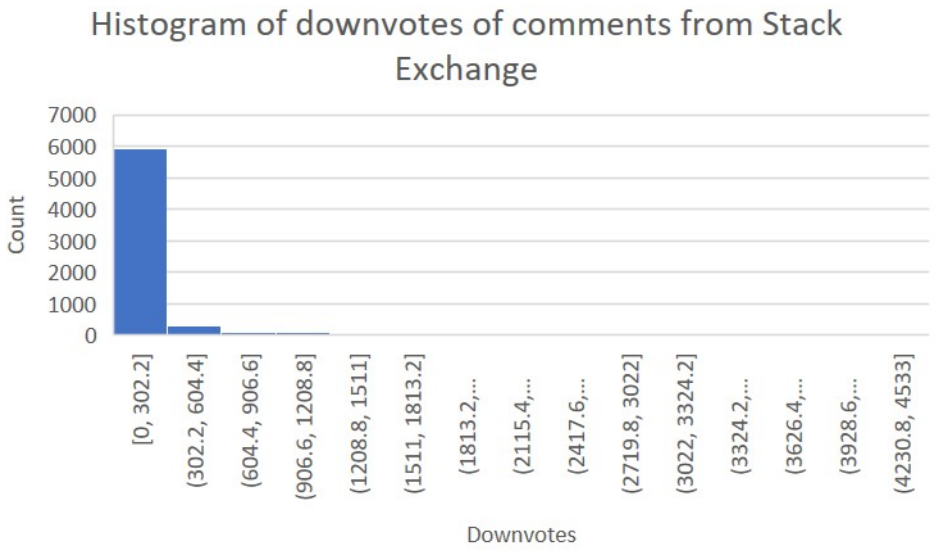

Figure 5.7: Histogram of number of downvotes on all comments

Table 5.4 and 5.5 show the descriptive statistics for the upvotes and downvotes variables.

Statistic	Value
mean	853.3854393
median	263
mode	68
variance	2220970.532
standard deviation	1490.292096

Table 5.4: Descriptive statistics for upvotes on comments

Statistic	Value
mean	108.5960882
median	15
mode	0
variance	89203.76603
standard deviation	298.6699952

Table 5.5: Descriptive statistics for downvotes on comments

Figure 5.8 shows the scatter plot between politeness score and upvotes on comments, from the Figure we cannot see any apparent correlation between these two variables, even the trend line in the figure does not show any significant relationship.

Figure 5.9 shows the scatter plot between politeness score and downvotes on comments, from the Figure we cannot see any apparent correlation between these two variables, even the trend line in the figure does not show any significant relationship. However, there is a slight downward trend in both the scatter plots. This downward trend indicates a slight negative correlation.

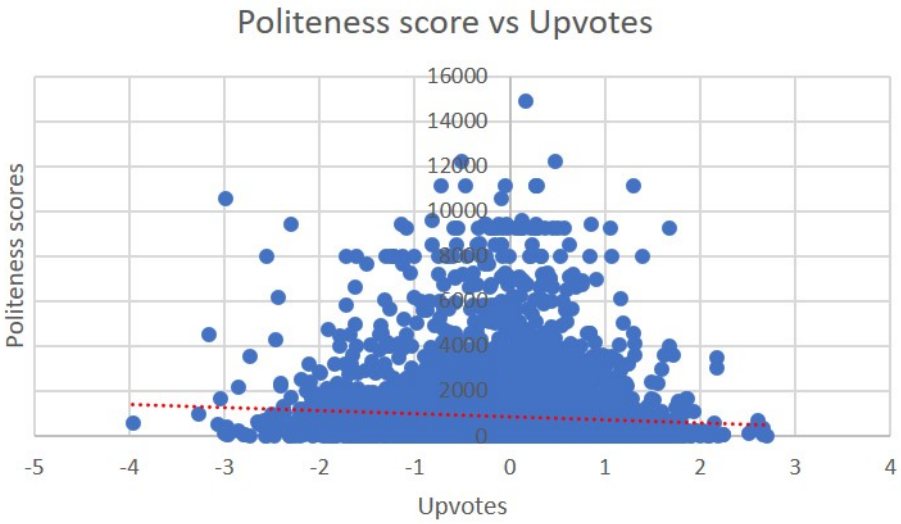

Figure 5.8: Scatter plot between politeness score of the comment and the upvotes it has

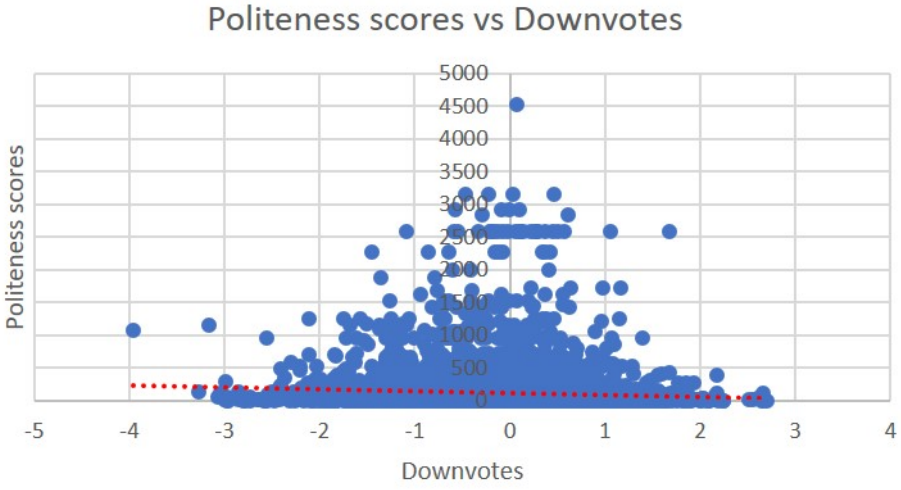

Figure 5.9: Scatter plot between politeness score of the comment and the downvotes it has

We have used Pearson's r correlation on the two variables. Person's r correlation is explained in Section 3.4. A Pearson's r correlation was calculated between the politeness score of the comment and the upvotes it gets, also a Pearson's r correlation was calculated between the politeness score of the comment and the downvotes it gets. We set our significance level, α, as 0.05. The results of our analysis an the Pearson's correlation are explained in Section 6.1.2. We have rejected the hypothesis based on the results from our analysis.

5.2 Using python PRAW package for collecting Reddit data

In section 3.1 we have a brief background of PRAW python package. This section elaborates on how this package was used for collecting data and what issues were encountered during data collection. Section 3.1 briefs user of the PRAW package in python and in this section we discuss the data collection strategy and issues encountered while collecting data.

We collected data from 4 subreddit forums ('Linguistics', 'Python', 'Words', 'WorldNews'). For collecting the data we had to obtain Reddit API credentials (Client ID, Client Secret, Username, Password and User Agent) from Reddit. Our data collection strategy was designed by keeping in mind the hypothesis we are trying to test. We collected comments which had a minimum of one reply and a maximum of 5 replies. This was done in order to avoid comments with no replies and comments with too many replies. There were issues with the PRAW package, if the user comment or reply had been deleted by the user then the package would give an error and the data collection script would brake down. This was overcome by ignoring these deleted comments and replies by using exception handling via try and except blocks.

5.3 Training model for predicting politeness scores using CRAN politeness package

Section 3.2 briefs user abut the CRAN package in R. The politenessProjection function in the package is used to train a LASSO regression model with Stack Exchange manually annotated comments dataset to predict the politeness scores of Reddit comments and replies dataset. The same model is also used to predict politeness of comments in rateme subreddit gender statistics dataset. LASSO regrresion is explained in Section 3.3. In the work of [9], authors use a similar methodology and predicts politeness scores of Wikipedia edit requests, they state that the model has a near human accuracy of predicting politeness. Figure 5.10 gives a snapshot of data which is fed in the model, not all columns are shown in the figure. In Figure 5.10 each column represents a politeness marker, each row represents a comment and the numerical data in the snapshot show represent occurrence of each politeness marker in the comments.

	Hedges	Positive.Emotion	Negative.Emotion	Impersonal.Pronoun	Swearing	Negation	Filler.Pause	Informal.Title	Formal.Title	Subjunctive	Indicative	By.The.Way	Let.Me.Know	Goodbye	For.Me	For.You
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	3	1	1	0	1	0	1	0	0	0	0	0	0	0	0
3	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0
5	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0
6	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	2	0	6	0	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0
11	2	4	0	3	0	3	1	0	0	0	0	0	0	0	0	0
12	0	3	2	3	0	2	0	0	0	0	0	0	0	0	0	0
13	0	3	0	2	0	0	0	0	0	0	0	0	0	0	0	0
14	2	3	1	2	0	0	0	1	0	0	0	0	0	0	0	0
15	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

Figure 5.10: Snapshot of data fed into the LASSO regression model

5.4 Analysis on Reddit Comments, Replies and Karma Data and their predicted politeness scores

The main goal of data collection from Reddit was to test our hypothesis. Section 4.2 explains the data collection strategy used to collect the data from Reddit, Figure 4.6 shows the snapshot of the dataset we collected from Reddit. Section 5.4.1 summarises the analysis done for the hypothesis testing.

5.4.1 Analysis on Hypothesis 3

The hypothesis we are trying to test here states, while interacting with similar posts on Reddit socially powerful people generally get more polite replies to their comments when compared to replies on comments of people with less social power. Figure 5.11 shows the methodology used for testing our hypothesis.

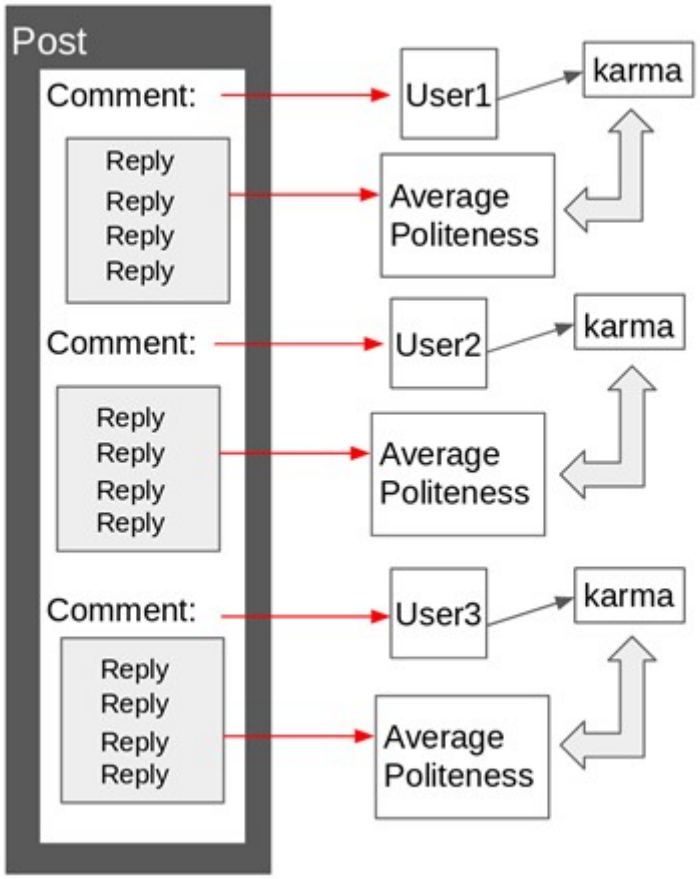

Figure 5.11: Methodology used for testing Hypothesis 3

For each reply, we calculate its politeness score using the R package proposed in [29] and explained in Sections 3.2 and 5.3. We average the politeness score of all replies for a comment. Finally, we use this average politeness score of replies and user information (karma, gender etc.) to test our hypothesis. Figure 5.12 shows the snapshot of data after collection from Reddit and predicting the politeness of replies. Figure 5.13 shows the snapshot of the data after our analysis.

S.No.	CommentNumber	Comment	karma	Replies	predictions
0	0	This is going to be use	377	people who ke	0.034354633
1	0	This is going to be use	377	Mind = blown	0.111583128
2	1	An alternative to this	527	PyOxidiser doe	-0.058228459
3	1	An alternative to this	527	Thanks to that	0.813590487
4	2	I failed on matplotlib.	5	I had a ton of	0.019213518
5	3	There's a gui that	32	Yes, but it dep	0.179829666

Figure 5.12: Snapshot of data after collection from Reddit and predicting the politeness score of replies

CommentNumber	AvgPolitenessScoreOfReplies	UserKarma
0	0.072968881	377
1	0.377681014	527
2	0.019213518	5
3	0.179829666	32
4	0.137063642	33
5	0.185740689	68

Figure 5.13: Reddit comments, replies and karma dataset after analysis

Figure 5.14 shows the histogram of the average politeness score of all replies for each comment in 4 different subreddits (linguistics, words, python, and world news). From the histograms in Figure 5.14 we can see that distribution of all these politeness scores is normal and all of them peak at politeness scores very near to 0. This also intuitively tells that our model is not predicting bizarre values.

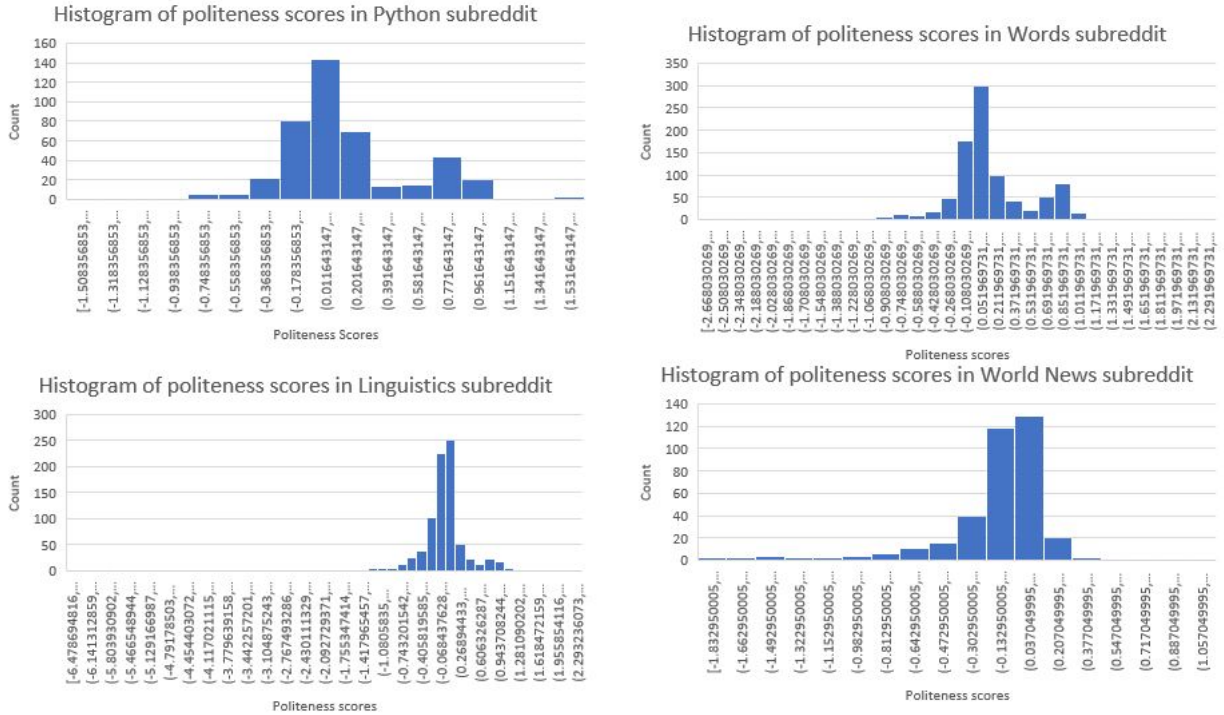

Figure 5.14: Histograms of average politeness score of replies for each comment in 4 different subreddits

Figure 5.15 shows the scatter plots between average politeness scores of replies for comments and the karma of the author of the comment to whom the replies are made to in 4 different subreddits.

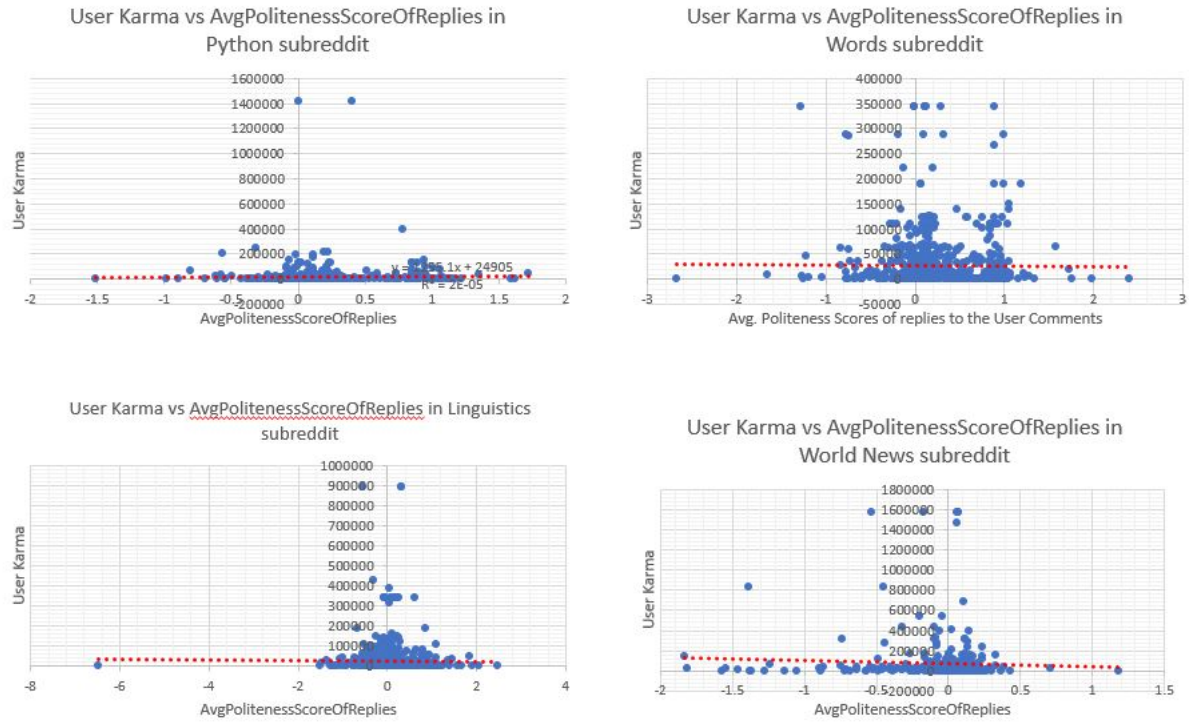

Figure 5.15: Scatter plot between of the average politeness score of replies for each comment and the use karma of the author of the comment in 4 different subreddits

From Figure 5.15 we can see no apparent relationship between these two variables in all the subreddits analysed. Trend line in the 4 plots also do not show any significant upward or downward trend. However, there is a slight downward trend in all the 4 scatter plots.

We have used Pearson's r correlation on the two variables. Person's r correlation is explained in Section 3.4. A Pearson's r correlation was calculated between the average politeness score of replies to comments and the user karma of users who posted the comments. We set our significance level, α, as 0.05. The results of our analysis and the Pearson's r correlation are explained in Section 6.1.3. We have rejected the hypothesis based on the results from our analysis.

5.5 Analysis on preprocessed rateme subreddit gender statistics dataset and their predicted politeness scores

In Section 4.3 preprocessing of rateme subreddit gender statistics dataset has been explained in detail. In this section we explain how this preprocessed dataset has been analysed to perform hypothesis testing. 4.7 shows the preprocessed dataset on which hypothesis testing is conducted. We tested three

hypotheses using this dataset. Section 5.5.1 explains our analysis on the hypothesis that the politeness of a comment is based on the gender of the user posting it: in particular, females are hypothesized to write more polite comments. Section 5.5.2 briefs the reader about our analysis on the hypothesis that politeness of a reply made to a comment is based on the gender of the person who made the comment: in particular, females get more polite replies to their comments. Section 5.5.3 walks reader through our analysis on the hypothesis that replies to the comments of the opposite gender are more polite when compared with replies to the comments of the same gender.

5.5.1 Analysis on Hypothesis 4

The hypothesis we are trying to test here states that the politeness of a comment is based on the gender of the user posting it: in particular, females are hypothesized to write more polite comments. Figure 5.16 shows the histograms of politeness scores of comments posted by males and females.

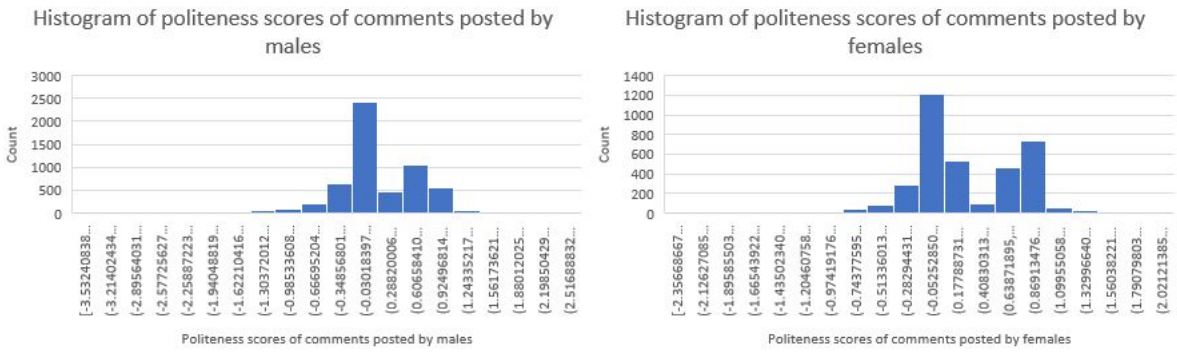

Figure 5.16: Histogram of politeness scores of comments posted by both the genders

From the histograms we can see that both the histogram peak near a value of 0, we can also see that in both the plots we have one other peak suggesting that mostly we have two types are users in this subreddit who are very polite and who are moderately polite. However, overall both the histograms are approximately normal. Since, both the histograms are normally distributed we can employ t tests for testing our hypothesis.

Figure 5.17 shows the box plot of the politeness scores of comments posted by both the genders. From further examination we can see that both mean and median of politeness scores of comments posted by females are higher than that of males. Table 5.6 shows the mean and median of politeness scores of comments posted by both the genders.

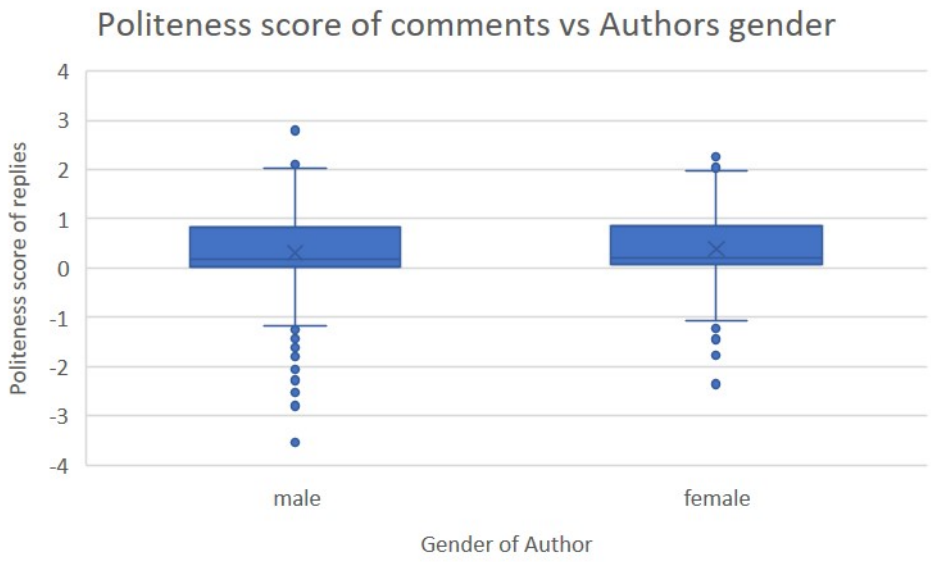

Figure 5.17: Box plots of politeness scores of comments posted by both the genders

Statistic	Males	Females
mean	0.304115472	0.386296813
median	0.1846	0.2156
variance	0.275197481	0.227781733

Table 5.6: Mean, median and variance for politeness scores of comments posted by both the genders

From table 5.6 we can see that mean of politeness scores of comments posted by females is higher than that of males. To check whether this difference is statistically significant we need to employ t-test. T-test is explained in Section 3.5. From Table 5.6 we can also see that the variances of the two variables differ. We employ a F-test to see whether there is a statistical difference in the variances of the two variables. F-test is also explained in Section 3.5. Our null hypothesis is that the variance of the two variables (politeness scores of comments posted by males and politeness scores of comments posted by females) are equal and our significance level is 0.05. After employing the F-test, We get a p-value of $2.46X(10)^{-10}$, which is less than our significance level, therefore we reject our null hypothesis. Keeping the result of F-test in mind we employ a two-tailed t-test assuming unequal variances with a null hypothesis that the means of the politeness scores of comments posted by males and politeness scores of comments posted by females are equal. Results of our analysis and the t-test are discussed in Section 6.1.4. Based on results of the t-test we reject the null hypothesis of the t-test that the means of the two variables are equal.

5.5.2 Analysis on Hypothesis 5

The hypothesis we are trying to test here states that politeness of a reply made to a comment is based on the gender of the person who posted the comment: in particular, females get more polite replies to their comments. Figure 5.18 shows the histograms of politeness scores of replies made to comments posted by males and females. For testing this hypothesis only comments with replies are considered.

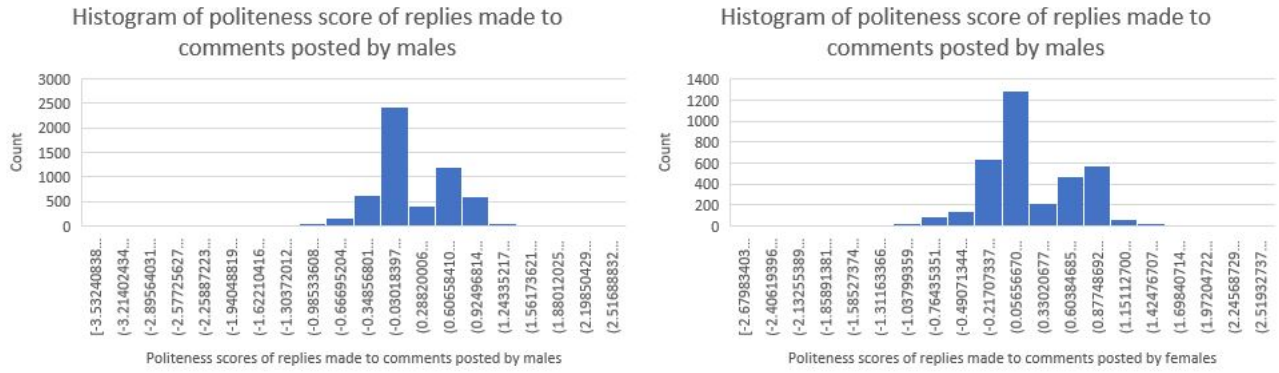

Figure 5.18: Histograms of politeness scores of replies made to comments posted by both the genders

From the histograms we can see that both the histogram peak near a value of 0, we can also see that, similar to our analysis in previous hypothesis, in both the plots we have one other peak suggesting that mostly we have two types are users in this subreddit who are very polite and who are moderately polite. However, overall both the histograms are approximately normal. Since, both the histograms are normally distributed we can employ t tests for testing our hypothesis.
Figure 5.19 shows the box plot of the politeness scores of comments posted by both the genders. From further examination we can see that both mean and median of politeness scores of replies made to comments posted by females and males are almost equal. Table 5.7 shows the mean and median of politeness scores of replies made to comments posted by both the genders.

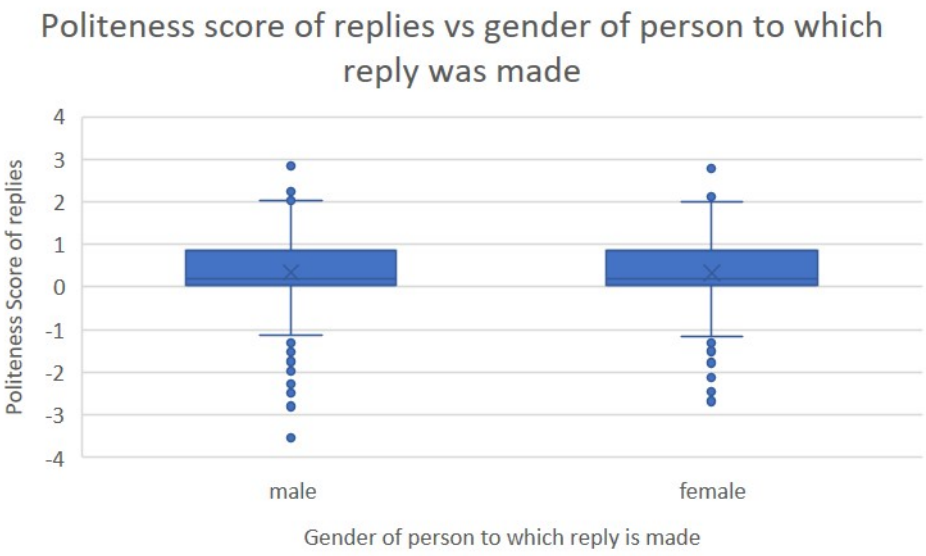

Figure 5.19: Box plots of politeness scores of replies made to comments posted by both the genders

Statistic	Males	Females
mean	0.338595566	0.332192946
median	0.199	0.199
variance	0.259769719	0.255917779

Table 5.7: Mean, median and variance for politeness scores of replies made to comments posted by both the genders

From table 5.7 we can see that there is no significant difference between mean of politeness scores of replies made to comments posted by females and mean of politeness scores of replies made to comments posted by males. To check whether the difference is statistically significant or not, we need to employ t-test. T-test is explained in Section 3.5. From Table 5.6 we can also see that the variances of the two variables do not differ significantly. We employ a F-test to see whether there is a statistical difference in the variances of the two variables. F-test is also explained in Section 3.5. Our null hypothesis is that the variance of the two variables (politeness scores of replies made to comments posted by males and politeness scores of replies made to comments posted by females) are equal and our significance level is 0.05. After employing the F-test, We get a p-value of **0.312**, which is higher than our significance level, therefore we fail to reject our null hypothesis.

Keeping the result of F-test in mind we employ a two-tailed t-test assuming equal variances with a null hypothesis that the means of the politeness scores of replies made to comments posted by males and politeness scores of replies made to comments posted by females are equal. Results of our analysis

and the t-test are discussed in Section 6.1.5. Based on results of the t-test we fail to reject the null hypothesis of the t-test that the means of the two variables are equal.

5.5.3 Analysis on Hypothesis 6

The hypothesis we are trying to test here states that replies to comments of opposite gender are more polite when compared with replies to comments of same gender. Figure 5.20 shows the histograms of politeness scores of replies made to comments posted by users of same and opposite gender. For testing this hypothesis only comments with replies are considered.

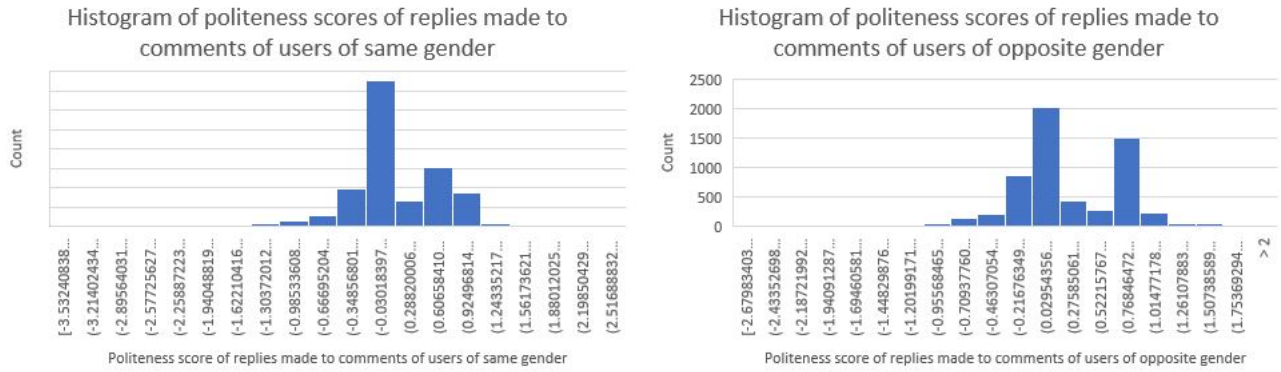

Figure 5.20: Histograms of politeness scores of replies made to comments posted by users of same and opposite gender

From the histograms we can see that both the histogram peak near a value of 0, we can also see that, similar to our analysis in previous hypothesis, in both the plots we have one other peak suggesting that mostly we have two types are users in this subreddit who are very polite and who are moderately polite. However, overall both the histograms are approximately normal. Since, both the histograms are normally distributed we can employ t tests for testing our hypothesis.

Figure 5.21 shows the box plot of the politeness scores of comments posted by users of same and opposite gender. From further examination we can see that both mean and median of politeness scores of replies made to comments posted by opposite gender are higher than that of mean and median of politeness scores of replies made to comments posted by same gender. Table 5.8 shows the mean and median of politeness scores of replies made to comments posted by users of same and opposite genders.

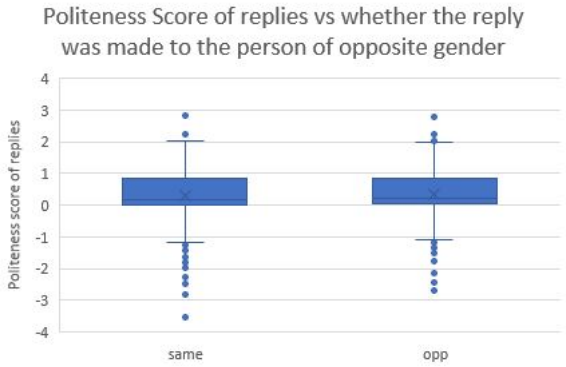

Figure 5.21: Box plots of politeness scores of replies made to comments posted by users of same and opposite genders

Statistic	same gender	opposite gender
mean	0.296006682	0.359368436
median	0.174	0.211
variance	0.276162963	0.246448766

Table 5.8: Mean, median and variance for politeness scores of replies made to comments posted by user of same gender and opposite gender

From table 5.7 we can see that there is a significant difference between mean of politeness scores of replies made to comments posted by user of same gender and mean of politeness scores of replies made to comments posted by user of opposite gender. To check whether the difference is statistically significant we need to employ t-test. T-test is explained in Section 3.5. From Table 5.6 we can also see that the variances of the two variables differ significantly. We employ a F-test to see whether there is a statistical difference in the variances of the two variables. F-test is also explained in Section 3.5. Our null hypothesis is that the variance of the two variables (politeness scores of replies made to comments posted by users of same gender and politeness scores of replies made to comments posted by users of opposite gender) are equal and our significance level is 0.05. After employing the F-test, We get a p-value of $8.60681X(10)^{-05}$, which is lower than our significance level, therefore we reject our null hypothesis.

Keeping the result of F-test in mind we employ a two-tailed t-test assuming unequal variances with a null hypothesis that the means of the politeness scores of replies made to comments posted by users of same gender and politeness scores of replies made to comments posted by users of opposite gender

are equal. Results of our analysis and the t-test are discussed in Section 6.1.6. Based on results of the t-test we reject the null hypothesis of the t-test that the means of the two variables are equal.

Chapter 6

Results and Discussion

In Chapter 5 we get a lot of interesting results. This chapter we state theses results and discuss their interpretations. Section 6.1 states these results elaborately and states based on these results whether we have successfully proven our initial hypotheses or not. Section 6.2 discusses more on these results and their interpretations.

6.1 Results

Results on each of our hypothesis are stated in Sections 6.1.1, 6.1.2, 6.1.3, 6.1.4, 6.1.4, 6.1.5, and 6.1.6.

6.1.1 Hypothesis 1

Our hypothesis is that the politeness of a comment is related to the reputation of the user who posted the comment. In Section 5.1.1 we explain that we have employed Pearson's r correlation on the two variables of interest, manually annotated politeness score of a comment and reputation of the author of the comment. The correlation value which we get is -0.08182, as stated in Section 5.1.1, we set our significance level, α, as 0.05. We need to calculate the critical value for seeing if our correlation is statistically significant or not. The concept of critical values has been explained in Section 3.4. The critical value for our correlation test with null hypothesis that there is no correlation between the two variables, given our significance level (0.05) and degrees of freedom (number of observations minus 2) is 0.024. As, magnitude of the correlation between two variables (0.08) is greater than the critical value we can reject the null hypothesis that the there is no statistical correlation between our variables.
Based on the hypothesis testing we can say that there is significant correlation between politeness score of a comment and the reputation of the user that posts it. However, the value of correlation is not that high that we can conclude that these two variables depend on each other entirely. There might be other factors which impose these correlations, some of these plausible factors are discussed in the

Section 6.2.

6.1.2 Hypothesis 2

Our hypothesis is that the politeness of a comment is related to the number of upvotes or downvotes it gets. In Section 5.1.2 we explain that we have employed Pearson's r correlation between manually annotated politeness score of a comment and number of upvotes it gets and we also Pearson's correlation between manually annotated politeness score of a comment and number of downvotes it gets. The correlation value between politeness scores and upvotes is -0.06763 and the correlation value between politeness scores and downvotes is -0.07585. As stated in Section 5.1.2, we set our significance level, α, as 0.05. We need to calculate the critical value for seeing if our correlation is statistically significant or not. The concept of critical values has been explained in Section 3.4. The critical value for our correlation test with null hypothesis that there is no correlation between the two variables, given our significance level (0.05) and degrees of freedom (number of observations minus 2) is 0.024. As magnitude of the correlation between politeness scores and upvotes (0.067) is greater than the critical value we can reject the null hypothesis that the there is not statistical correlation between these two variables. Also, magnitude of the correlation between politeness scores and downvotes (0.075) is greater than the critical value we can reject the null hypothesis that the there is no statistical correlation between these two variables.

Based on the hypothesis testing we can say that there is significant correlation between politeness score of a comment and the number of upvotes and downvotes it gets. However, the value of correlation is not that high that we can conclude that these two variables depend on each other entirely. There might be other factors which impose these correlations, some of these plausible factors are discussed in the Section 6.2.

6.1.3 Hypothesis 3

Our hypothesis is that while interacting with similar posts on Reddit socially powerful people generally get more polite replies to their comments when compared to replies on comments of people with less social power. In Section 5.4.1, we explain that we have employed Pearson's r correlation between the average politeness score of replies made to comments and the user karma of users who posted the comments to which the replies were made to in 4 different subreddits. Table 6.1 shows the correlation value between these two variables in the 4 different subreddits. As stated in Section 5.1.2, we set our significance level, α, as s0.05. We need to calculate the critical value for seeing if our correlation is statistically significant or not. The concept of critical values has been explained in Section 3.4. The critical value for our correlation test with null hypothesis that there is no correlation between the two

variables, given our significance level (**0.05**) and degrees of freedom (number of observations minus 2) are given in Table 6.1.

Subreddit	Pearson's r correlation	Critical value
Linguistics	-0.0116	0.069
Python	0.004954953	0.096
Words	-0.008790871	0.068
World News	-0.044056238	0.107

Table 6.1: Correlation values and critical values in 4 different subreddits

As magnitude of the correlation between the average politeness score of replies made to comments and the user karma of users who post the comments is less than the critical value in all the subreddits, hence, we fail to reject the null hypothesis that the there is no statistical correlation between these two variables.

Based on the hypothesis testing we can say that we fail to find any significant correlation between the average politeness score of replies made to comments and the user karma of users who post the comments. Section 6.2 discusses more on these results.

6.1.4 Hypothesis 4

Our hypothesis is that the politeness of a comment is based on the gender of the user posting it: in particular, females are hypothesized to write more polite comments. In Section 5.5.1 we explain that we have employed t-test to see whether there is a statistical difference between the mean of politeness score of comments posted by females and mean of politeness score of comments posted by males. As stated in Section 5.5.1, we set our significance level as a p-value of **0.05**. If the p-value we get after employing t-test is less than **0.05**, then we reject the null hypothesis that the means are same, otherwise we fail to reject the null hypothesis of equal means. Section 3.5 explains hypothesis testing using t-test. For the t-test with the null hypothesis as, means of politeness scores of comments posted by males and females are same, we get the p-value of $\mathbf{8.85956}X\mathbf{(10)}^{-15}$. As, resulting p-value ($\mathbf{8.85956}X\mathbf{(10)}^{-15}$) is less than our significance level(p-value of **0.05**) so we can reject the null hypothesis that the the means of both the variables are equal.

Based on the hypothesis testing we can say that mean of politeness scores of comments posted by females is statistically different than mean of politeness scores of comments posted by males. From Table 5.6 we can see that mean of politeness score of comments posted by females are greater than that of males. So, we can say that females write more polite comments than males and politeness score

of the comment is related to the gender of the person who writes that comment. Section 6.2 discussed more about these results.

6.1.5 Hypothesis 5

Our hypothesis is that the politeness of a reply made to a comment is based on the gender of the person who posted the comment: in particular, females get more polite replies to their comments. In Section 5.5.2 we explain that we have employed t-test to see whether there is a statistical difference between the mean of politeness scores of replies made to comments which were posted by females and mean of politeness scores of replies made to comments which were posted by males. As stated in Section 5.5.2, we set our significance level as a p-value of **0.05**. If the p-value we get after employing t-test is less than **0.05**, then we reject the null hypothesis that the means are same, otherwise we fail to reject the null hypothesis of equal means. Section 3.5 explains hypothesis testing using t-test. For the t-test with the null hypothesis as, mean of politeness scores of replies made to comments which were posted by females is equal to the mean of politeness scores of replies made to comments which were posted by males, we get the p-value of **0.554910629**. As, resulting p-value (**0.554910629**) is greater than our significance level(p-value of **0.05**) so we fail to reject the null hypothesis that the the means of both the variables are equal.

Based on the hypothesis testing we fail to say that mean of politeness scores of replies made to comments which were posted by females is statistically different than mean of politeness scores of replies made to comments which were posted by males. Also, from Table 5.7 we can see that both the means do not differ significantly. So, we can say that we fail to prove our initial hypothesis that the politeness of a reply made to a comment is based on the gender of the person who posted the comment: in particular, females get more polite replies to their comments. Section 6.2 discussed more about these results.

6.1.6 Hypothesis 6

Our hypothesis is that replies to comments of opposite gender are more polite when compared with replies to comments of same gender. In Section 5.5.3 we explain that we have employed t-test to see whether there is a statistical difference between the means of the politeness scores of replies made to comments posted by users of same gender and politeness scores of replies made to comments posted by users of opposite gender. As stated in Section 5.5.3, we set our significance level as a p-value of **0.05**. If the p-value we get after employing t-test is greater than **0.05**, then we reject the null hypothesis that the means are same, otherwise we fail to reject the null hypothesis of equal means. Section 3.5 explains hypothesis testing using t-test. For the t-test with the null hypothesis as, mean of the politeness scores of replies made to comments posted by users of same gender is equal to mean of the politeness scores

of replies made to comments posted by users of opposite gender, we get the p-value of $1.21904X(10)^{-08}$. As, resulting p-value ($1.21904X(10)^{-08}$) is less than our significance level(p-value of 0.05) so we reject the null hypothesis that the the means of both the variables are equal.

Based on the hypothesis testing we can say that mean of the politeness scores of replies made to comments posted by users of same gender is statistically different than the mean of the politeness scores of replies made to comments posted by users of opposite gender. Also, from Table 5.8 we can see that the mean of the politeness scores of replies made to comments posted by users of opposite gender is higher than the mean of the politeness scores of replies made to comments posted by users of same gender. So, we can say that we prove our initial hypothesis that replies to comments of opposite gender are more polite when compared with replies to comments of same gender. Section 6.2 discussed more about these results.

6.2 Discussion

In Section 6.1 we see the results of our analyses and hypothesis testing. In this Section, we try and interpret these results and discuss about them. In Sections 6.1.1, 6.1.2 and 6.1.3, all most all of the correlation values are slightly negative (except for one with a correlation value of almost 0). Also, in Figures 5.5, 5.8, 5.9 and 5.15 we can observe a slight downward trend in the trend line. The plausible reasons behind this can be that we are examining social media forums which tend to be impolite. This impoliteness in social media forums could be because we don't personally know the person with whom we are interacting, also users have very limited access to the information about the person who they are interacting with. From Sections 6.1.1, 6.1.2, and 6.1.2 we can see that we get very low correlation values between politeness scores and the variables of interest (Userkarma, reputaion, upvotes and downvotes), however we hypothesised intuitively that these variables should have high correlation values because in face to face conversations these variables do impact the politeness of a conversation. This can be due to the fact that in face to face conversations we can get an idea about social power of the person we are interacting with by his personality, talking style, clothing sense etc. however, not all social media allows us to see the social power of the person we are interacting with. Most of the times, on Social media which gives information about user's social power through variables likes, upvotes or user karma, it is hard to spot the variable of social power in between a interaction or conversation. Therefore, plausible reasons behind seeing a lot of low and negative correlations in the results of hypothesis in Sections 6.1.1, 6.1.2, and 6.1.3 could be because of impoliteness nature of social media forums and unawareness of a user's social power in between interactions on these forums. Interaction plot Figure 6.1 summarise the results in Sections 6.1.4, 6.1.5, and 6.1.6. In Figure 6.1 y-axis indicate the politeness score of a reply, x axis indicate the gender of the person who made the reply,

and the pattern of lines indicate the gender of the person to which the reply was made to. In Figure 6.1 we can observe that when females reply to females they are less polite whereas when females reply to males they are more polite. A similar statement could be made about males, they reply to females more politely than males. Also, it is quite evident from the figure that overall replies written by females are more polite than replies written from males. The plausible reason behind these results could be that that gender of the user is visible to the other user's in most of the social media forums. Also, if the gender is not visible in the social forum then one could easily guess the gender of a person while interacting by intuitively analysing differences user's reaction to different interactions and differences in writing styles. The main plausible reason of the results in Sections 6.1.4, 6.1.5, and 6.1.6 could be availability and awareness of user's gender information on these forums.

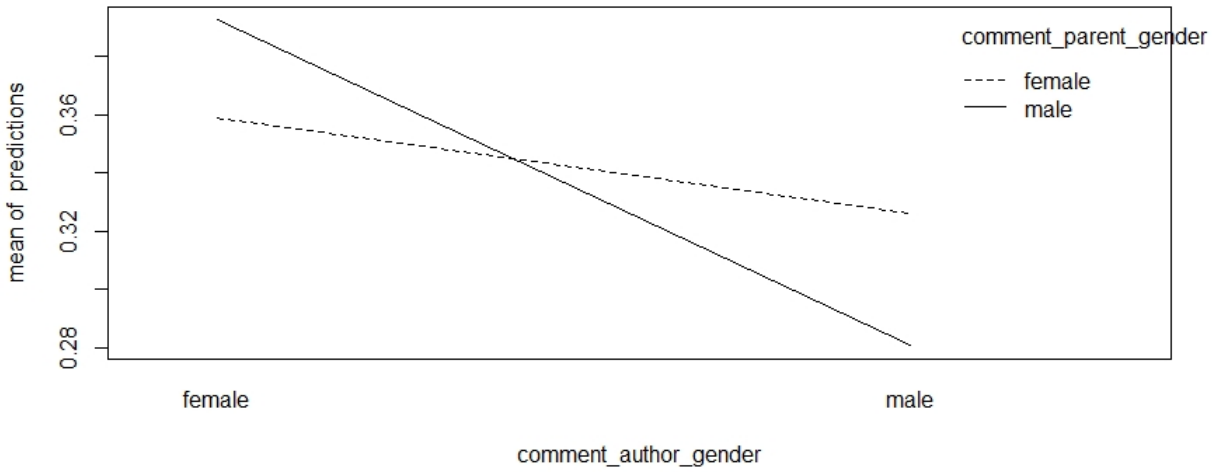

Figure 6.1: Interaction plot

Chapter 7

Conclusion and Future work

7.1 Future Work

The research tries to answer the research question as briefly as possible but there is always a room for further improvement and future work. Due to limited time, there were only a few hypotheses which were tested, however, this research can easily be extended to test other hypotheses based on the politeness theory. The hypotheses in this work only try to correlate two factors, Social power (reputation, upvotes, downvotes on Stack Overflow and Karma points on Reddit) and gender, with politeness theories nevertheless hypotheses based on factors like age can also be tested using the same methodology. Only a few social media forums are used in this study to perform hypothesis testing, it would be very interesting to see of the results of this study are true on other popular social media websites like Facebook, Twitter, Instagram, etc. The biggest limitation of this work is that it uses only one model for predicting politeness scores, comparison with other machine learning models could have been done in order to measure the performance of the model. Also, currently the research is limited to a manually annotated dataset on politeness to measure the performance of the model used, however, this might not be the case in the future when we have sufficient data to check our model performance. This study takes the first step in the direction of analyzing politeness theories on social media and leaves a lot of scope for extending this work in the future.

7.2 Conclusion

In this study, various hypotheses are formulated based on politeness theory. These hypotheses are then tested over social media to see whether they work or not. In order to perform hypothesis testing various tools, techniques and datasets are investigated. Along with data pre-processing of investigated datasets, data collection on a social media forum, Reddit, is also performed in order to test our hypotheses on befitting datasets. Using the investigated tools and techniques, a LASSO regression

model is trained over existing manually annotated datasets to predict politeness scores of the collected dataset. Results of the study state that on Reddit, comments posted by females are more polite when compared to comments posted by males, however, comments posted by females don't get more polite replies when compared to replies made to comments posted by males, nevertheless, replies made to comments posted by users of the opposite gender are more polite than replies made to comments posted by users of the same gender. From the results of this study, it can also be said that that there is a statistically significant correlation between politeness of a score and the reputation of a user posting the comment over Stack Exchange, however, the correlation score is not that high that this can be stated with confidence. The results also say that there is a statistically significant correlation between a comment's politeness score and the number of upvotes and downvotes it gets on Stack Exchange but again the correlation score is not that high that this can be said with high confidence. By analyzing 4 different subreddits the study also rejects the hypothesis that socially powerful people generally get more polite replies to their comments when compared to replies on comments of people with less social power. This research contributes to the existing literature on politeness theory and tries to bridge the current gap of analyzing politeness theory on social media. The results of this study show that there is a huge potential in carrying out research in the domain of politeness theory and its aspects over social media. The research tries to answer the research question in detail and leaves out room for further investigation.

Bibliography

[1] Malika Aubakirova and Mohit Bansal. Interpreting neural networks to improve politeness comprehension. arXiv preprint arXiv:1610.02683, 2016.

[2] Mansurul Bhuiyan. Tone Analyzer for Customer Engagement: 7 new tones to help you understand how your customers are feeling. https://www.ibm.com/blogs/cloud-archive/2017/04/tone-analyzer-customer-engagement-7-new-tones-help-understand-customers-feeling/, 2017. Accessed: 2018-10-31.

[3] Bryce Boe. Praw: The python reddit api wrapper. https://praw.readthedocs.io/en/latest/, 2015. Accessed: 2019-08-01.

[4] Penelope Brown, Stephen C Levinson, and Stephen C Levinson. Politeness: Some universals in language usage, volume 4. Cambridge university press, 1987.

[5] CRAN. The comprehensive r archive network. https://cran.r-project.org/. Accessed:2019-15-07.

[6] Jonathan Culpeper. Towards an anatomy of impoliteness. Journal of pragmatics, 25(3):349–367, 1996.

[7] Jonathan Culpeper. Impoliteness and entertainment in the television quiz show: The weakest link. Journal of Politeness Research. Language, Behaviour, Culture, 1(1):35–72, 2005.

[8] Cristian Danescu-Niculescu-Mizil. Wikipedia edit requests and stack exchange comments dataset. http://www.cs.cornell.edu/~cristian/Politeness_files/Stanford_politeness_corpus.zip, 2013.

[9] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. A computational approach to politeness with application to social factors. arXiv preprint arXiv:1306.6078, 2013.

[10] Penelope Eckert and Sally McConnell-Ginet. Language and gender. Cambridge University Press, 2013.

[11] Gino Eelen. A critique of politeness theory, volume 1. Routledge, 2014.

[12] Erving Goffman. Interaction ritual: Essays in face-to-face behavior. Routledge, 2017.

[13] H Paul Grice. Logic and conversation. william james lectures. Studies in the Way of Words, pages 1–138, 1967.

[14] H Paul Grice, Peter Cole, Jerry Morgan, et al. Logic and conversation. 1975, pages 41–58, 1975.

[15] Krishna Hariramani. Reddit Comment reply dataset. https://github.com/krishzinx/MastersThesis, 2019.

[16] Robin Lakoff. The logic of politeness; or minding your ps and qs papers from the 9th regional meeting of the chicago linguistic society. Chicago: Chicago Linguistic Society, pages 292–305, 1973.

[17] Alexander Lyzhov. Gender statistics of /r/RateMe. https://www.kaggle.com/nikkou/gender-statistics-of-rrateme, 2018.

[18] Sara Mills. Gender and politeness, volume 17. Cambridge University Press, 2003.

[19] Albert Park, Mike Conway, and Annie T Chen. Examining thematic similarity, difference, and membership in three online mental health communities from reddit: a text mining and visualization approach. Computers in human behavior, 78:98–112, 2018.

[20] K Pearson. Mathematical contributions to the theory of evolution. iii. regression. Heredity and Panmixia. Philosophical.

[21] John Priniski and Zachary Horne. Attitude change on reddit's change my view. In CogSci, 2018.

[22] PyPA. pip the python package installer. https://pip.pypa.io/en/stable/, 2008. Accessed: 2019-08-01.

[23] Ron Scollon, Suzanne Wong Scollon, and Rodney H Jones. Intercultural communication: A discourse approach. John Wiley & Sons, 2011.

[24] George W Snedecor and William G Cochran. Statistical methods, eight edition. Iowa state University press, Ames, Iowa, 1989.

[25] Student. The probable error of a mean. Biometrika, pages 1–25, 1908.

[26] Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288, 1996.

[27] Richard J Watts. Politeness. Cambridge University Press, 2003.

[28] Gregor Wiedemann. Opening up to big data: Computer-assisted analysis of textual data in social sciences. Historical Social Research/Historische Sozialforschung, pages 332–357, 2013.

[29] Michael Yeomans, Alejandro Kantor, and Dustin Tingley. The politeness package: Detecting politeness in natural language. R Journal, 10(2), 2018.