Title
What Did 60,000 People Write in Their Dating App Essays? Text Analyses Using Various Techniques

Author
Yang, Grace

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

What Did 60,000 People Write in Their Dating App Essays? Text Analyses Using Various Techniques

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics

by

Grace Huey Yang

2021

ABSTRACT OF THE THESIS

What Did 60,000 People Write
in Their Dating App Essays?
Text Analyses Using Various Techniques

by

Grace Huey Yang
Master of Applied Statistics
University of California, Los Angeles, 2021
Professor Ying Nian Wu, Chair

Abstract: A 2019 survey reported that roughly one-third of adults in the U.S. had at some point or other used a dating app or website. When we dissected almost 60,000 dating self-summary essays, what interesting insights emerged? Did text analysis on essays reveal any age differences? What about comparisons between gender, different education level, or cat- versus dog-lovers? Did the essays cluster into distinct groups when applying classification modeling techniques? Can each essay be mapped spatially and similar essays found this way? How can dating companies like Tinder and Match make use of these text analysis insights?

The thesis of Grace Huey Yang is approved.

Akram M. Almohalwas

Mark S. Handcock

Rick Schoenberg

Ying Nian Wu, Committee Chair

University of California, Los Angeles

2021

TABLE OF CONTENTS

1 Introduction **1**

1.1 Data Source 2

1.2 Text Analysis Scope 2

1.3 A Search for Similarities 3

2 Sentiment Analysis **5**

2.1 Lexicons Used 5

2.2 Sentiments in General 5

2.3 By Gender 7

2.4 By Age 8

2.5 By Education 10

2.6 By Religion 10

2.7 By Pet Preference 11

3 High-Dimensional Embeddings **13**

3.1 Embeddings 13

3.2 t-Distributed Stochastic Neighbor Embedding 14

3.3 Essay Length and Effect on Embeddings 15

3.4 Cluster Visualization by Age Group 18

3.5 Cluster Visualization by Gender 18

3.6 Cluster Visualization by Pet Preference 19

4 Approximate Nearest Neighbors **20**

4.1 Subject Essay 1 and Its Closest Essays 20
4.2 Subject Essay 2 and Its Closest Essays 22

5 Classification 25
5.1 Support Vector Clustering 25
5.2 Classification by Gender 26
5.3 Classification by Age Group 27

6 Correlation of Essays 29
6.1 By Age, by Gender, by Orientation 29
6.2 By Education, by Racial Group 30
6.3 By Religion 31
6.4 By Drink Habit, by Pet Preference 32

7 Topic Modeling 34
7.1 Latent Dirichet Allocation 34

8 Word Frequencies 35
8.1 Word Count by Age, by Gender 35
8.2 Most Frequent Words 36

9 Word Networks 38
9.1 Bi-grams 38
9.2 Word Networks 38

10 Most Unique Words 41
10.1 X-Y Plots of TF-IDF 41

10.2 By Gender 42

10.3 By Age Group 43

10.4 By Pet Preference 45

11 Business Applications 47

12 Conclusion 48

References 49

LIST OF FIGURES

1.1 Age distribution of dating singles 3

2.1 Most frequent positive and negative words using Bing lexicon 6
2.2 Emotions using NRC lexicon 6
2.3 Sentiment scores by gender 7
2.4 Most frequent positive and negative words by gender 8
2.5 Sentiment scores for women's essays 9
2.6 Sentiment scores for men's essays 9
2.7 Sentiment scores by education level using AFINN lexicon 10
2.8 Sentiment scores by religion using AFINN lexicon 11
2.9 Sentiment scores by pet preference using AFINN lexicon 12

3.1 Example of King - Man + Woman = Queen 14
3.2 t-SNE visualization for MNIST handwritten digits 0 to 9 15
3.3 Visualize embeddings of a short sentence versus complete essay 16
3.4 Visualize embedding clusters by age 18
3.5 Visualize embedding clusters by gender 19
3.6 Visualize embedding clusters by pet preference 19

5.1 Example of face classification using SVC 25
5.2 Heat map of face classification 26
5.3 Accuracy of gender classification 26
5.4 Heat map of gender classification 27

5.5 Accuracy of gender classification 27
5.6 Heat map of gender classification 28
5.7 Accuracy of gender classification 28

6.1 Essay correlations between age, gender, orientation 29
6.2 Essay correlations between education, racial groups 31
6.3 Essay correlations between religions 32
6.4 Essay correlations between drinking habits, pet preference 33

7.1 LDA topic modeling with k = 3 34

8.1 Length of dating profile essays by gender and age 36
8.2 Most common words in dating essays 37

9.1 "Love" bi-grams in women's essays 39
9.2 "Love" bi-grams in men's essays 39
9.3 Word networks in women's essays 40
9.4 Word networks in men's essays 40

10.1 Side-by-side view of word dispersion as age gap increases 42
10.2 Compare word frequencies by gender 42
10.3 Compare word frequencies of age 20s vs 30s 43
10.4 Compare word frequencies of age 20s vs 40s 44
10.5 Compare word frequencies of age 20s vs 50s 45
10.6 Compare word frequencies by pet preference 45

LIST OF TABLES

3.1 Varying character lengths of an essay 17

CHAPTER 1

Introduction

Online dating is one of the most popular ways that people use to find romantic relationships in the 21st century. In a 2019 survey conducted by the Pew Research Center, nearly 50% of young Americans said they had used dating websites and/or mobile apps (Pew, 2020). The experience had been largely favorable, with six out of 10 online daters reporting positive feedback about using these dating platforms to find romantic partners and love.

Besides asking that daters provide the usual information about gender, age, sexual orientation, height, geographical location, etc., the online dating website and apps usually also ask that the daters write up some personal essays. These essays are likely meant to help flesh out a bit more about the individual's likes and dislikes, their personalities and what their lifestyles are. If filled out with honesty and thought by the individuals, these free-form self-summaries may be able to provide additional insight and more qualitative understanding into a person's more holistic, inner self.

The range and depth of text analysis research and applications have really exploded in the past decade. For example, we are now able to apply machine learning to detect the patterns that define spam emails and filter them out, train business chat-bots that can respond in a human-like way to customers' most frequently asked questions, build recommendation engines that find the closest matching news or social media content to what was just read or viewed by people.

It begs the question: what can we uncover from dating essays with text data analysis?

1.1 Data Source

There is a public data set of nearly 60,000 dating profiles from OkCupid, a free dating app and website (Kim and Escobedo-Land, 2015). This data set was scraped by researchers in mid-2012 and represented the people who were active on OkCupid in the last 12 months ending June 2012.

There were over 30 features for each dater in this data set:

- Demographic: age, education, ethnicity, gender, income, religion
- Physical: body type, height
- Habits: diet, drug use, alcohol use, tobacco use, orientation, pets
- Personal essays: 10 sections with the first being a self-summary
- Other: job, location, offspring status and preference, relationship status, languages spoken and fluency, horoscope sign, date last online on OkCupid app

There were more men than women in this data set: split roughly 60/40 men to women. Figure 1.1 shows the age distribution of the individuals. The mean age of the dating singles was 32-33 years old; the median age was 30.

1.2 Text Analysis Scope

I performed text analyses on the self-summary essays. The features of interest were: age, gender, sexual orientation, education level, religion, pet preference.

The techniques applied to the data included: high-dimensional text embeddings, t-distributed stochastic neighbor embedding and cluster visualization, Support Vector Clustering and Classification, Approximate Nearest Neighbors of embeddings in space, sentiment

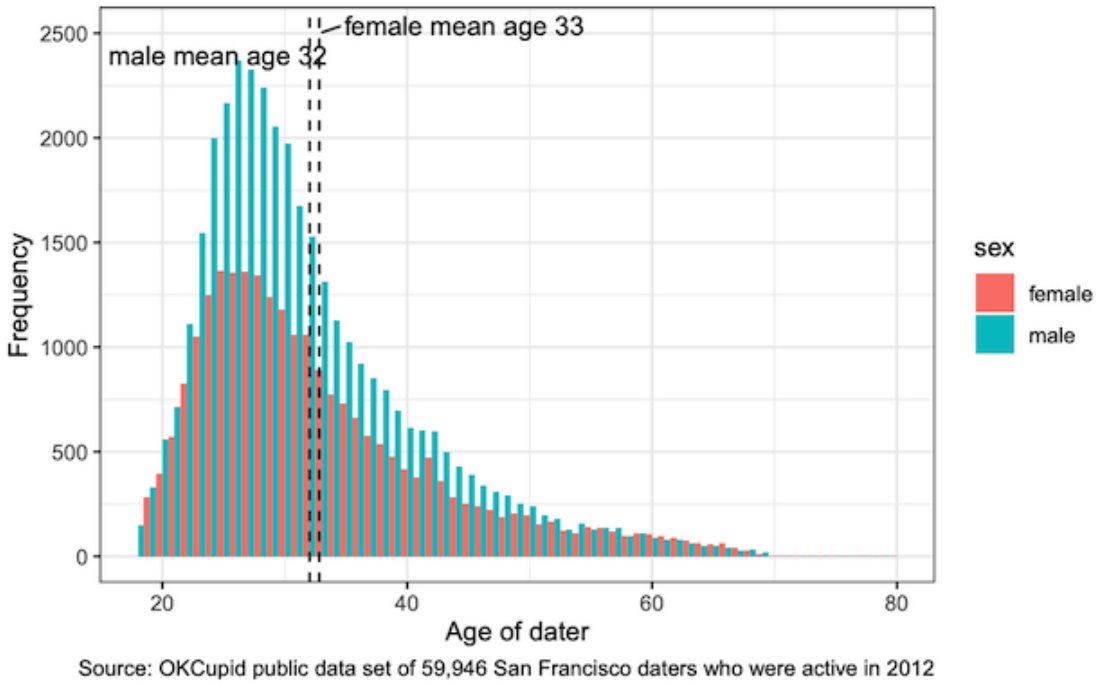


Figure 1.1: Age distribution of dating singles

analysis of the dating essay corpus, word networks, and Latent Dirichlet Allocation topic modeling.

While this was far from an exhaustive coverage into all text analyses methods and all 30+ available features, it should still be a good representation of text-mining techniques and patterns found from dating essays.

1.3 A Search for Similarities

Text analyses on what people write in their dating profile self-summaries may help uncover similarities or differences for different features like gender, age group, socio-economic status, etc.

While there are certainly many cases of opposites attracting, it is more common for people to gravitate towards others of similar age, educational background, general intelligence,

physical attractiveness level, religious faith, interests, habits, manner of speech, etc. (Buss, 1985)

When behavioral researchers studied the personality types of Facebook users and their social network on the app, they found that the social network—friends and romantic partners—tended to have personalities similar to the individual (Youyou et al., 2017). Other research in this area has also consistently found links between personality and language use (Hirsh and Peterson, 2009; Argamon et al., 2005; Tausczik and Pennebaker, 2010). Extroverts, for example, are more likely to use words like "bar", "shots", "Miami", "pool", "restaurant", and "dancing" than more introverted people (Yarkoni, 2010).

My focus was to find similarities between the dating essays using the following text analysis approaches:

- Comparisons: I explored what similarities or differences existed in the way that people presented themselves on dating app essays. There were both quantitative and qualitative measurements, such as whether the essay sentiments skewed towards the positive or negative, the essay lengths, the word choices made in the essays, etc.

- Embeddings: I transformed the dating essays were into high-dimensional, information-rich embeddings. Each essay embedding can be thought of as being a unique vector in space, surrounded by other essay vectors, some closer and others further away. I found approximate nearest neighbors for any individual essay based on cosine distances between embedding pairs. This has business use in recommendation systems.

- Classification: I fitted classification models to the dating essays. The models probed whether the words and context had patterns that predicted features such as an individual's gender, age, orientation, etc. The high-dimensional embeddings were also reduced down to 2-dimension space to visualize and find close-matching clusters.

CHAPTER 2

Sentiment Analysis

2.1 Lexicons Used

I applied three popular lexicons or dictionaries used for sentiment analysis on the dating essay corpus:

- "Bing" (Hu and Liu, 2004) where words are classified into either the positive or negative category, and
- "AFINN" (Nielsen, 2011) which scores words going from -5 to 5 with the negative scores representing negative sentiments and positive scores indicating positivity.
- "NRC" Emotion Lexicon (Mohammad and Turney, 2013) which organizes words into one of eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, or disgust) or one of two polarities (negative or positive).

2.2 Sentiments in General

Both AFINN and NRC lexicons scored the dating essays as being on the positive spectrum. About 80% of the essays were scored as being positive by the NRC lexicon. The average AFINN score was 1.45.

The top three essay words that were categorized as of positive sentiments are love, like and good, as shown in Figure 2.1. This made intuitive sense: the dating websites and

apps prompt customers to share something about their hobbies ("I love music and movies") and habits ("I like running and hiking", "I enjoy good food and backyard barbeques with friends") in hopes of attracting like-minded others.

Figure 2.1: Most frequent positive and negative words using Bing lexicon

In terms of word choices in the essay corpus, the NRC lexicon categorized many into emotions of joy, trust, and anticipation (Figure 2.2).

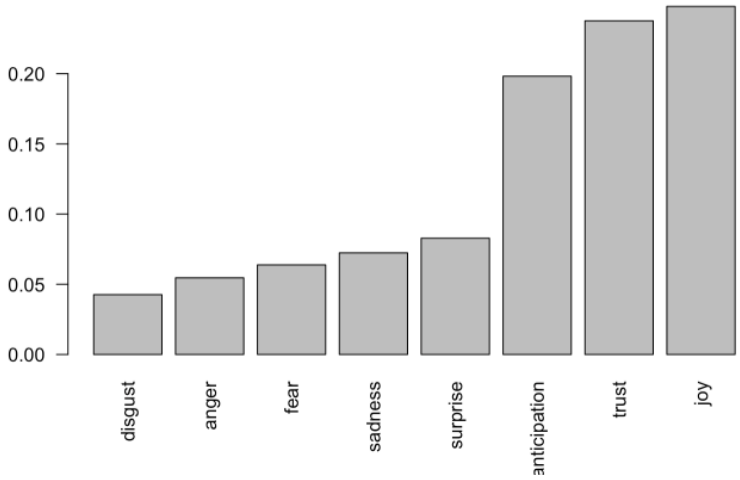


Figure 2.2: Emotions using NRC lexicon

2.3 By Gender

Women's essays generally scored slightly more positive than men's on the AFINN lexicon's -5 to +5 score range: 1.52 versus 1.4 (Figure 2.3).

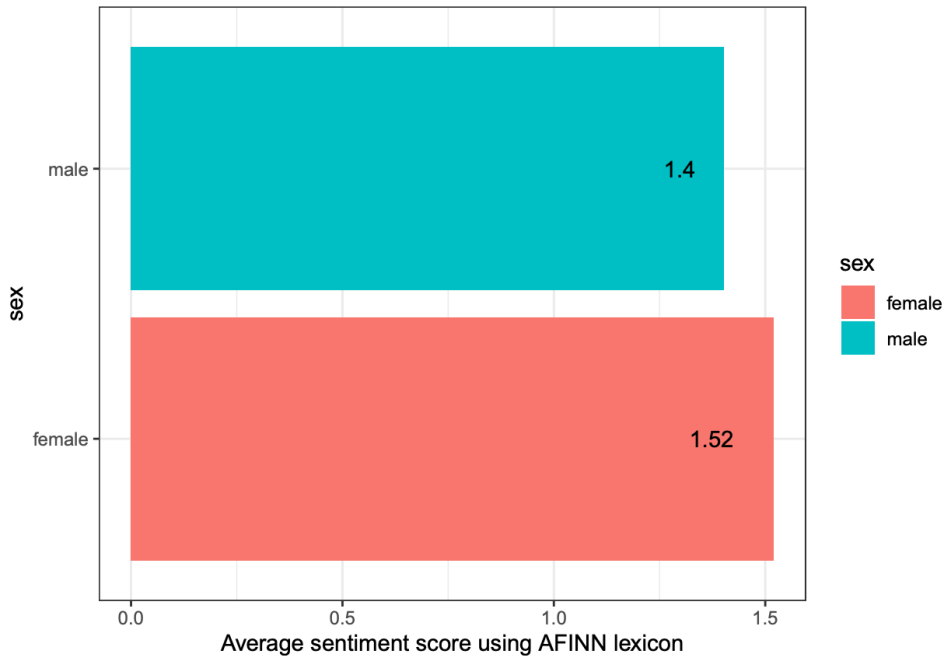
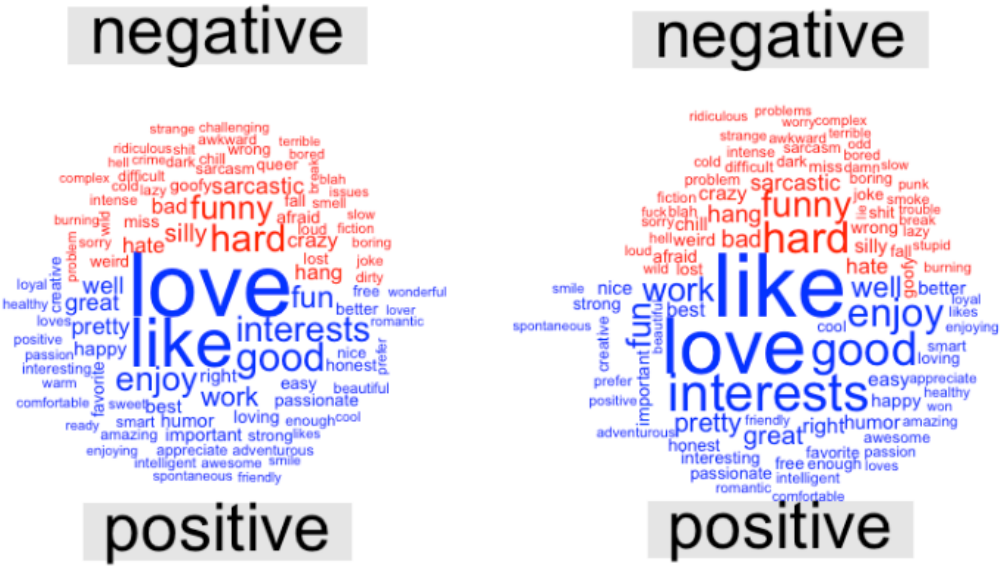


Figure 2.3: Sentiment scores by gender

Comparing word clouds of the most frequently occurring terms side-by-side (Figure 2.4), we did notice that dating essays for men and women did overlap in the top positive words "love" and "like" from the Bing lexicon.

Looking more closely at these word clouds by gender, it appeared that there was a difference in the word order of the top two. In women's essays (Figure 2.4a), "love" dominated and was the top word followed by the next most frequent word "like". Men's essays showed that "like" came in as the most common word before "love" (Figure 2.4b).

There are societal-shaped gender differences in most countries. Male gender "standards" lean more toward being strong and emotionally stoic. Using the word "love" has cultural



(a) Women's essays (b) Men's essays

Figure 2.4: Most frequent positive and negative words by gender

connotations of softness and femininity. Women are generally allowed more emotional leeway and tend to have less reticence about using the emotionally-elevated word "love".

The dating essays in this data set revealed that cultural gender norms were well and alive.

2.4 By Age

There was also an improving trend in sentiment scores from the AFINN lexicon for the older age groups (Figure **??**). It was observed to a modest degree when analyzing men's essays. For women's essays, the increase in sentiment scores toward the more positive was more pronounced than the men's.

The AFINN average score for female daters' essays in their twenties was 1.4 while those in their fifties and older scored 1.75 on average (Figure 2.5). The improvement was 0.35 for

women over an age gap of 30 years.

Compare this to the average sentiment score for men's essays: 1.33 for those in their twenties and 1.53 for the fifties and older (Figure 2.6). The sentiment score increase was 0.2 for men over the same 30-year age gap.

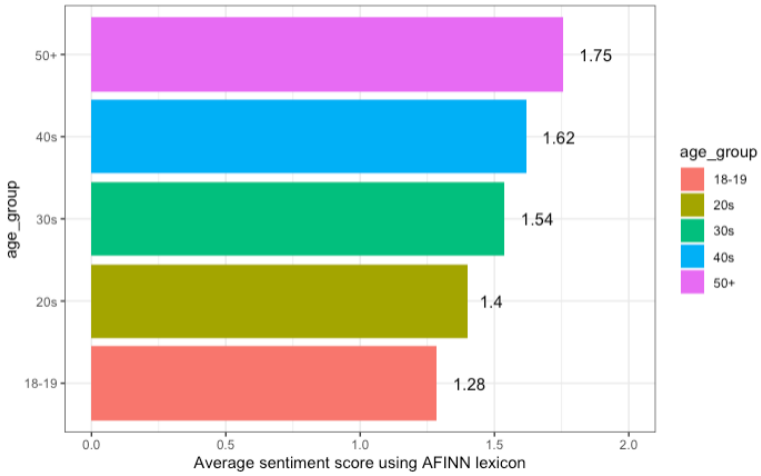


Figure 2.5: Sentiment scores for women's essays

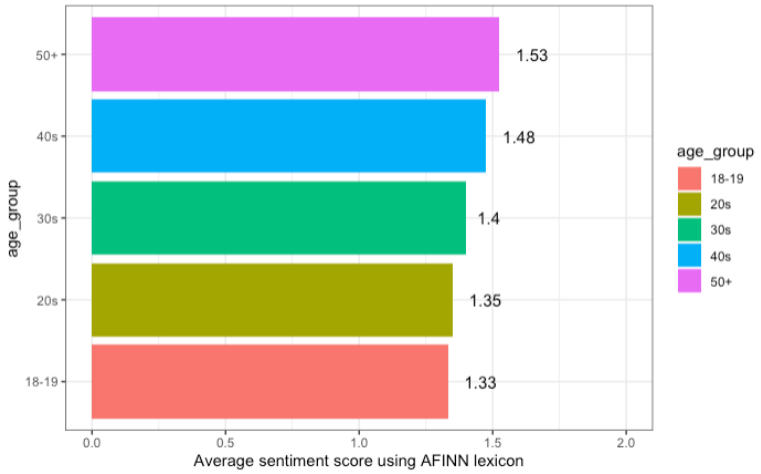


Figure 2.6: Sentiment scores for men's essays

2.5 By Education

There were also differences in dating essay sentiment scores (AFINN lexicon) depending on the level of education completed by individuals (Figure 2.7). Those who had post-graduate degrees scored 1.54 in essay sentiment compared to 1.35 for high school graduates. Dating essays of bachelor degree holders scored in the middle of the pack: 1.47 on average.

Higher scores on the AFINN -5 to +5 scale indicate that text is more positive in sentiments.

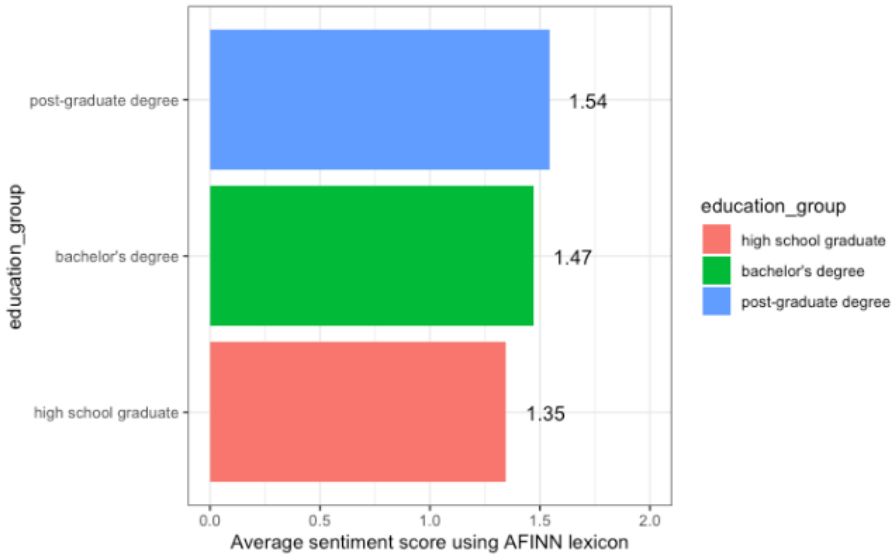


Figure 2.7: Sentiment scores by education level using AFINN lexicon

2.6 By Religion

Let us turn our attention now to sentiment scores by major religious groups. The average scores for each group using the AFINN lexicon were summarized in Figure 2.8.

We saw Christianity, Catholicism and Judaism having the highest sentimental essay scores out of the different faith groups. The dating singles in this data set who had indicated that they were atheists had the lowest average sentiment scores relative to other major

religious groups. The gap between atheism and Christianity, Catholicism was roughly 0.3 points.

Dating essays for the agnostic group scored closer to the atheist group than to the Christian and Jewish faith groups. Buddhism scored in the middle range of all faith groups.

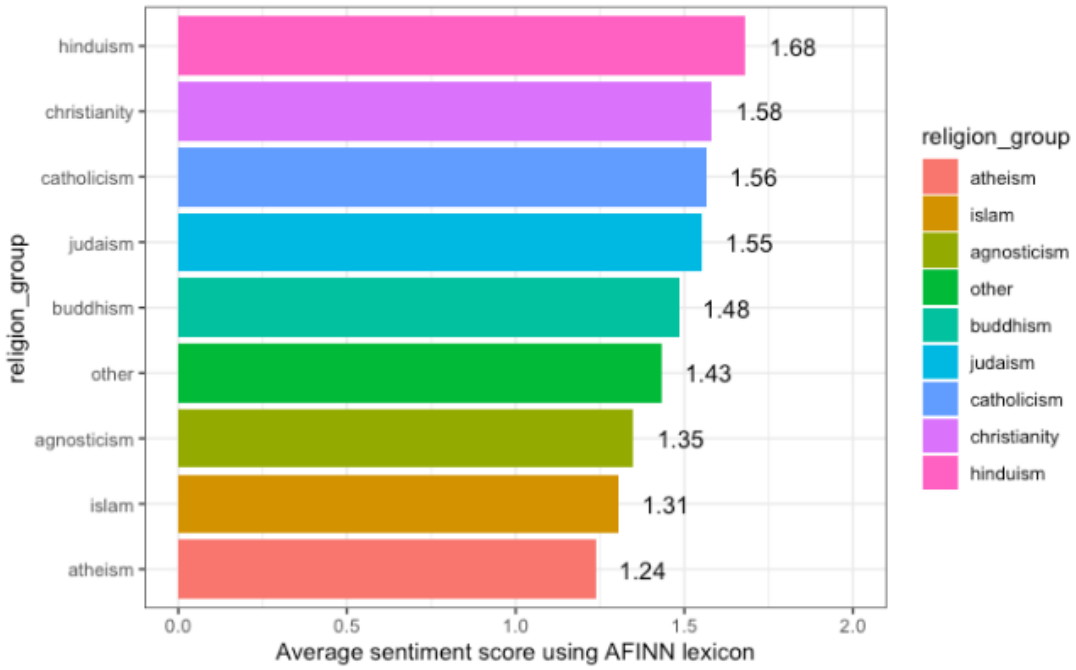


Figure 2.8: Sentiment scores by religion using AFINN lexicon

2.7 By Pet Preference

It was interesting to compare sentiment scores along the spectrum of cat versus dog lovers. The OkCupid app allowed users to indicate a range of pet animals (dogs, cats) and pet preferences (whether the individual has, likes, or dislikes one or both animal types). The OkCupid user could also choose not to respond to this optional pet question.

The dating essays were first grouped into three broad categories: likes cats, likes dogs, and likes both cats and dogs before text sentiment analysis.

Dog lovers appeared to show a small lead over cat lovers: sentiment scores were 1.52 on the positive AFINN scale for dog aficionados compared to 1.33 for the pro-cat crowd (Figure 2.9). Dog owners in this data set might have tended to be more extroverted than cat owners. Extroversion can be observed in how people express themselves, both in spoken words and written prose.

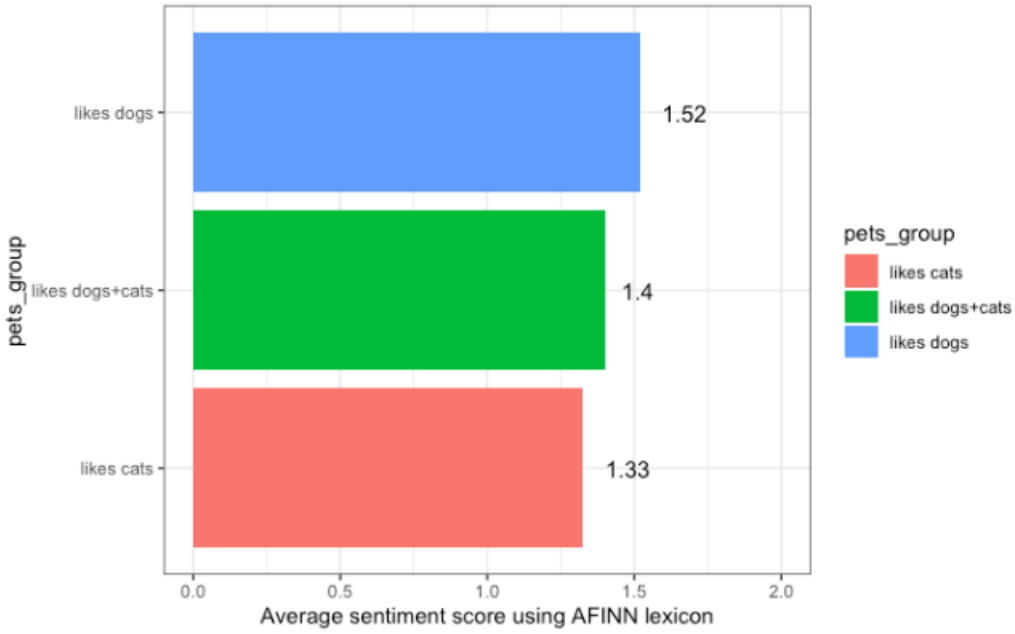


Figure 2.9: Sentiment scores by pet preference using AFINN lexicon

CHAPTER 3

High-Dimensional Embeddings

3.1 Embeddings

"Hey Google, navigate to the Griffith Observatory."

Natural Language Processing models are the keystone to enabling computers to learn human languages and respond similarly. We use this more and more in our daily activities, for example asking Google or Siri for a stock market update, weather report, top news stories. Not to be outdone, technology behemoth Amazon also has its version Alexa, which can play music on-demand, put in a grocery order at Whole Foods, and turn your smart home devices on or off with voice commands.

The first steps involve machine learning on text data, after which the text gets transformed into high-dimensional numerical vectors or embeddings. At present, the number of resulting dimensions can range from hundreds to thousands.

The high-dimensionality makes a lot of sense when we think back on the incredible richness of human language. After the text data gets converted into numbers and vector space, we can pass it on down the machine learning pipeline. Neural networks have made significant inroads into learning the meaning of inidividual words, sequences of words, and context.

There is an often-quoted example to illustrate the power of language embeddings uses the three words "king", "man", and "woman". First, we take the three individual word embeddings, then apply simple vector arithmetic of this formula: "king" - "man" + "woman".

Then we compare it to the embeddings of all other words in the English language. The new embedding is found to be closest to that for the word "queen" (Figure 3.1).

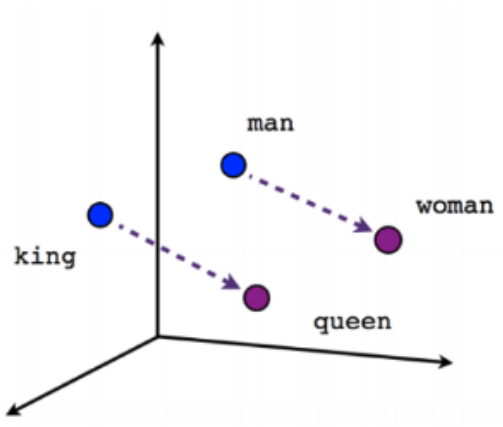


Figure 3.1: Example of King - Man + Woman = Queen

This result is amazing if we try to think about how machine learning had been able to come up with intelligent word embeddings that correctly depicted the meanings and word relationships.

3.2 t-Distributed Stochastic Neighbor Embedding

While it is easy to understand and visualize two or three dimensions, it becomes impossible with the high-dimensional space of text embeddings. Google and Facebook have a couple of the current state-of-the-art language embedding algorithms out there. Google's Universal Sentence Encoder which produces 512 embedding dimensions while Facebook's Infersent churns out a whopping 4,096 dimensions!

One of the ways of trying to visualize high-dimensional data is a tool called t-distributed Stochastic Neighbor Embedding or t-SNE. With t-SNE, we can look at the embeddings in two or three dimensional space. This is useful for checking if clusters existed in the data.

A well-known example is the application of t-SNE to the MNIST database of handwritten

digits from 0 to 9. Each digit's image is 28 by 28 pixels, or 2828 = 784 dimensional vector. t-SNE does a fairly good job at getting out the hidden patterns in the nearly 800 dimensions and tranform them down into just two "condensed" dimensions for visualization purposes (Figure 3.2).

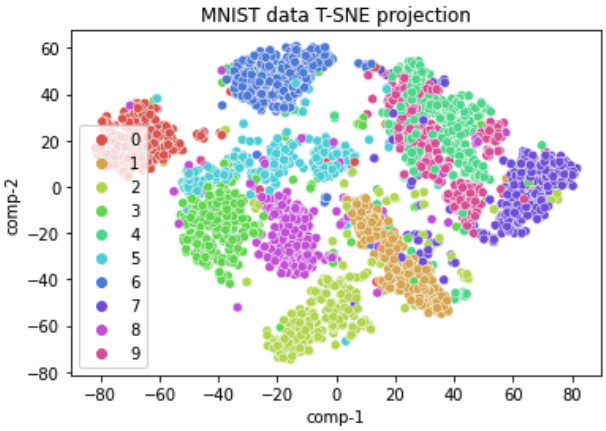


Figure 3.2: t-SNE visualization for MNIST handwritten digits 0 to 9

There are ten relatively well-defined clusters, one for each of the digits "0" to "9". It is important to note though that distinctly isolated visual clusters are not a guarantee; how well the visual clusters are observed with t-SNE is predicated on clusters existing in the data.

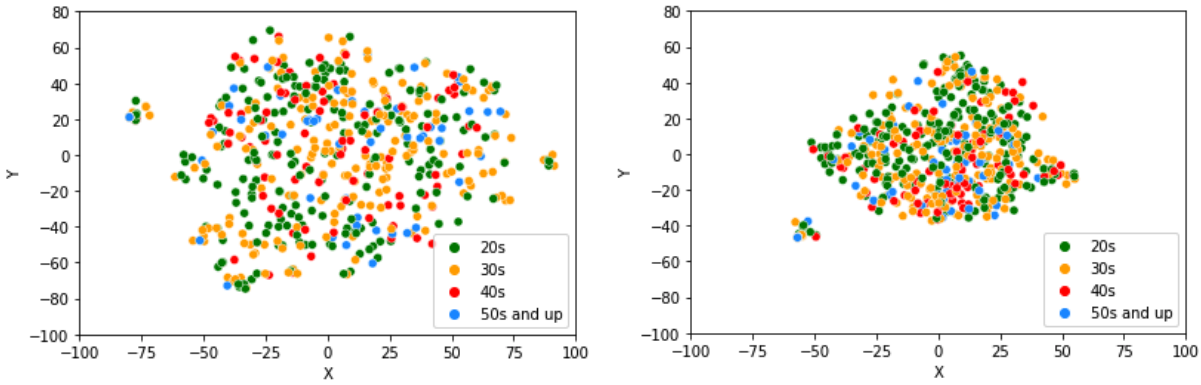
3.3 Essay Length and Effect on Embeddings

High-dimensional text embeddings on single sentences (news headlines) and short paragraphs (book or movie synopses) are highly useful in today's internet industry. They are essential components to recommendation engines used by Google, Netflix, and Hulu. In terms of how much we feed into the embedding machine, was more always better? Or was there a strong case to be made for choosing brevity instead?

Data science practitioners should remain cognizant of the effects of increasingly long text

imput. Google had this to say about its embedding engine Universal Sentence Encoder: "... Universal Sentence Encoder embeddings also support short paragraphs. There is no hard limit on how long the paragraph is. Roughly, the longer the more 'diluted' the embedding will be..."

What was the quality of "diluted" represented in a t-SNE reduced, two-dimensional embedding space? To help answer this question, we examined a side-by-side comparison when using the first 30 characters of each essay to create the embeddings (Figure 3.3a) versus using the complete versions (Figure 3.3b). We saw that when short essay snippets were used, embeddings were more scattered compared to the tighter center mass observed for complete essays.



(a) First 30 characters per essay

(b) Complete essays

Figure 3.3: Visualize embeddings of a short sentence versus complete essay

The ability to see clusters (if any) appeared to favor shorter lengths. But how short should data scientists go in the world of dating essays? Are tweet-lengths on Twitter app a good benchmark for the 21st century, internet-era human's attention span?

Let us first look at a sample essay that is 221 characters long:

"Well-traveled, over-educated, but so far under-appreciated. I grew up in Santa Monica, did my undergrad at Berkeley, got my MA in London, and then

returned to the Bay Area for law school. Now an attorney in San Francisco."

Next we examined snippets of various character lengths (Table 3.1) and the information provided at each length:

Number of characters	Essay to get embeddings for
First 30 characters	"Well-traveled, over-educated, "
First 70 characters	"Well-traveled, over-educated, but so far under-appreciated. I grew up "
First 100 characters	"Well-traveled, over-educated, but so far under-appreciated. I grew up in Santa Monica, did my underg"
First 140 characters	"Well-traveled, over-educated, but so far under-appreciated. I grew up in Santa Monica, did my undergrad at Berkeley, got my MA in London, an"
All 221 characters	"Well-traveled, over-educated, but so far under-appreciated. I grew up in Santa Monica, did my undergrad at Berkeley, got my MA in London, and then returned to the Bay Area for law school. Now an attorney in San Francisco."

Table 3.1: Varying character lengths of an essay

We made some general inferences below. This helped us narrow down suitable dating essay lengths for deriving text embeddings:

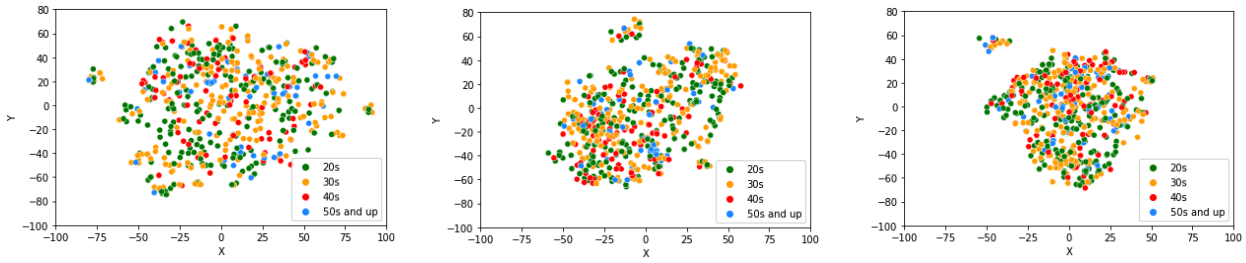
- 30 characters appeared to have provided a decent start to knowing who this individual was and what made them tick.
- Between 70 to 100 characters, we learned that this person was probably of higher socio-economic level since he/she wrote well, used good grammar and spelling, and showed a level of wit. 70 to 100 characters is the length of good "tweets" that were liked and most often retweeted on the Twitter app.

- At 140 characters (or the character count limit for Twitter up through late-2017), this individual signaled that he/she was of a high intelligence level through name-dropping a great school and that he/she had received many advanced degrees.
- Other personality facets which were revealed with the complete essay: this person was a world citizen who had lived in many large cosmopolitan cities, and his/her current profession is being a lawyer.

3.4 Cluster Visualization by Age Group

In this section, we explored whether any well-defined, meaningful embedding clusters existed with character lengths 30, 70 and 140.

Figure 3.4 showed the progression of going from 30 to 70 to 140 characters. There were not really clear clusters that could be visualized.



(a) First 30 essay characters (b) First 70 essay characters (c) First 140 essay characters

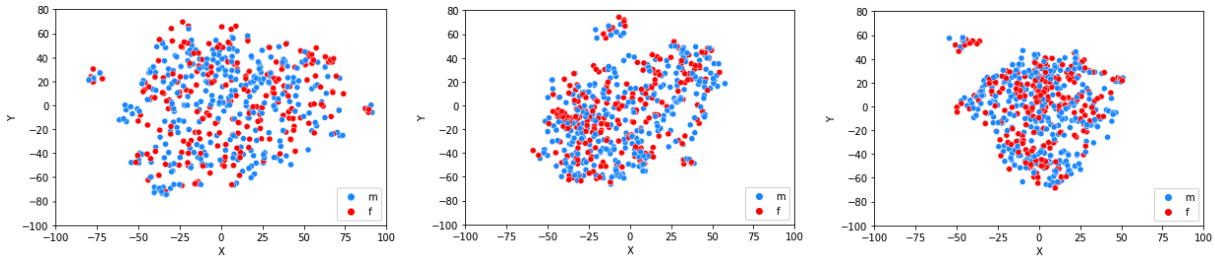
Figure 3.4: Visualize embedding clusters by age

3.5 Cluster Visualization by Gender

Were there patterns in women's essays that separate them from the men's?

The essay embeddings after passing them through t-SNE are shown in Figure 3.5. There were not well-defined, meaningful embedding clusters that could be visualized this way. This

was the case for different essay snippets of the first 30, 70, and 140 characters.



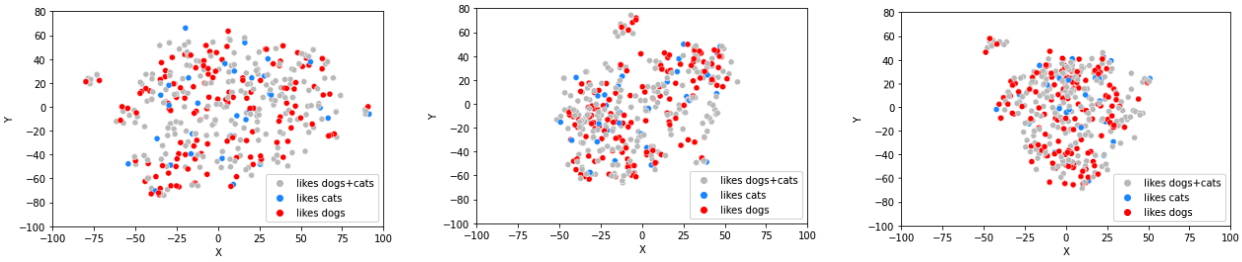
(a) First 30 essay characters (b) First 70 essay characters (c) First 140 essay characters

Figure 3.5: Visualize embedding clusters by gender

3.6 Cluster Visualization by Pet Preference

Were there any well-defined, meaningful embedding clusters existed for people's pet preferences? Did we see cat-lovers in one clear cluster and dog-lovers in another?

No distinctly isolated clusters were observed in the t-SNE transformed essay embeddings (Figure 3.6). This was true for the shortest to longest character lengths of dating essays.



(a) First 30 essay characters (b) First 70 essay characters (c) First 140 essay characters

Figure 3.6: Visualize embedding clusters by pet preference

CHAPTER 4

Approximate Nearest Neighbors

For popular dating apps and websites like Tinder, Bumble, Match and OKCupid, their customer base ranges between two to eight million people. These companies employ a variety of filtering, ranking and search algorithms in order to make match recommendations to their users.

It is seldom done with brute computation force across all possible pairings or combinations. Let us use an example of a dating app with a customer base of four million users, split 50/50 between men and women. If the dating company was interested in searching for the closest matching essay embeddings for its users aged 20 to 29 years old, this narrows the field down to say one million men and one million women. The number of possible pair-wise comparisons that would have to be computed is one million times one million, or over one trillion!

Instead a common technique used that makes embedding comparisons more efficient is called Approximate Nearest Neighbors or ANN. Given a customer's self-summary essay or "subject essay" and the position of its embedding in the high-dimensional space, ANN will search for other essay whose embeddings are mapped closer to this subject essay.

4.1 Subject Essay 1 and Its Closest Essays

We selected two daters' essays from the OkCupid data set and put them through the ANN search algorithm.

This is subject essay 1:

"Well-traveled, over-educated, but so far under-appreciated. I grew up in Santa Monica, did my undergrad at Berkeley, got my MA in London, and then returned to the Bay Area for law school. Now an attorney in San Francisco."

Top ANN essays to subject essay 1 (the first three as an example):

- "I live on the Peninsula (Belmont). Worked as an engineering manager in Silicon Valley. Now retired and widowed. Two grown kids and 2 grandsons. Education: B.S. in Engineering from UC Berkeley. Certificate in Management from Univ of Santa Clara. I have traveled extensively in Europe, Asia, and the Middle East. My primary hobby is fly fishing, and I have traveled to Canada, Mexico, Alaska, the Bahamas to do it and even fished in Japan."
- "Hi, I'm from Santa Cruz, CA. I moved to San Francisco about five years ago to go to college, and I spent one of those years studying abroad in Japan! Graduated last year and just landed a job. I commute to Silicon Valley for work. I'm still getting used to the city. I'd like to explore more of SF! You could call me a Japanophile, I like Japan and learning Japanese! It's hard but fun! And I like traveling in Japan and hopefully other countries in the future."
- "I grew up in Cincinnati, OH and moved to the Bay Area 7 years ago. I love the San Francisco and consider this city my home."

Looking at the top three closest-matching essays, the common threads to subject essay 1 included: a few moves around the country, appeared to have a love for international travel, advanced educational degrees. The high-dimensional embeddings seemingly did a good job of capturing the dating essays.

4.2 Subject Essay 2 and Its Closest Essays

This is subject essay 2:

> *"My name's Jake. I'm a creative guy and I look for the same in others. I'm easy going, practical and I don't have many hang ups. I appreciate life and try to live it to the fullest. I'm sober and have been for the past few years. I love music and I play guitar. I like tons of different bands. I'm an artist and I love to paint/draw etc. and I love creative people. I've got to say I'm not too big on internet dating. you cant really get an earnest impression of anyone from a few polished paragraphs. but we'll see, you never know."*

The following closest matches to subject essay 2 were found using the ANN technique (first three as an example):

- "I will give this a shot, but I'm no poet! I'm a positive person who likes to be around people. I try to find the best in everybody. I have a huge heart, I'm genuine, caring, and kind. I don't take things too seriously. you can usually find me laughing about something. I love my family and I enjoy spending time with my parents (maybe because I'm an only child?). I'm actually first generation since both of my parents are Czech immigrants (they have an interesting story). I'm from the bay area, went to UCLA for college, and then moved to Prague for a few years and now I'm back where I started. I found a career in the video game industry and I feel very fortunate for many reasons. I love spending time with my friends and I'm lucky to work with amazing people that are like my second family. I really enjoy traveling. I love exploring new places and immersing myself in foreign cultures. Otherwise, I am happy if I can find time to oil paint and practice yoga. I love trying new restaurants and making recommendations to friends (spread the love). No, seriously, food and wine bring me great pleasure. I enjoy going to art museums, comedy shows, and concerts, yadda yadda yadda. I like to get

outside to do some hiking, skiing, and walking around our beautiful city! Chemistry is everything so none of this matters anyway..haha"

- "I'm Italian. I'm discovering the world, living in different places between Italy and unite states making a cool job, very exiting and challenging. I like discover things and be free to do what I like to do. I love learning and speak with new people to know their stories and feeling differences and similitudes between me and them. Open-minded .. Sincere (and expect my partner to be as well)... Dreamer (but hard worker to make dreams real) .. Happy about my life (and ready to change if something doesn't satisfy me anymore) .. Driven by curiosity, looking for challenges. I love travel and I'm interested in different cultures and ways of living. I love sushi and sashimi, Thai, Indian, Mexican, and of course Italian. I love the good wine but I'm little ignorant about it! I also love listen music but I'm not that kind of guy that know everything about it. I used to have a motorcycles but after an incident I decided to be more calm and use only the car or the bicycle! ;-) I have lived in San Francisco, Turin, Milan, Rimini, and for few time also in Monaco, London, Shanghai, Rome, ect.. I'd like to do a lot of things... maybe too much! I'm not a difficult person, but I don't like losing my time. I like watching movies and making good conversations I'm looking for a nice girl, dynamic, intellectual, patient, and sweet, with small and big dreams and the passion to follow them. I'm looking for a partner to share all these experiences! I want to have fun but I'm also ready for something important. Honestly I don't like stay here and have spent so much time to think and write this words but you know that find cool people around it's difficult ... We are all busy for 1000 things and work hard... So let's try, maybe you also have the same feeling, and we need to pass trough this site to know each other :-)"

- "I'm a very friendly, caring, and open-minded person. People fascinate me, so I'm often inclined to ask a lot of questions (it's a good thing; it just means I'm interested). I love to laugh and somehow, I find that I'm often attracted to witty and sarcastic

human beings. I love exploring and learning new things. I have an endless bucket list of things I would like to accomplish before I die, and I am most excited to do so. I work hard and I play hard. I think it's important to find that balance. In a relationship, I value honesty, respect, and communication, but most importantly, I long for romance and passion. I believe a relationship is reciprocal and both parties should get the same thing they put into it."

For subject essay 2 and the top three closest essays, all seemed to have easy-going personalities, creative, right-brained, straight-forward, down-to-earth types. This is another anecdotal example of personalities and traits being well-represented by the embeddings.

CHAPTER 5

Classification

5.1 Support Vector Clustering

Support Vector Machines (SVM) and Supported Vector Clustering (SVC) are some of the frequently used tools for the purposes of classification. Both are very powerful tools, with applications in both image and text classifications.

Figure 5.1 shows an example of 24 images and names identified by SVC. 23 out of 24 of the face samples were correctly classified with people; only one of George W. Bush was incorrectly identified as the former U.K. Prime Minister Tony Blair. If the classification was perfect, the F1 score would be value of one. The lowest possible F1 value is zero, which happens if either the precision or the recall is zero.

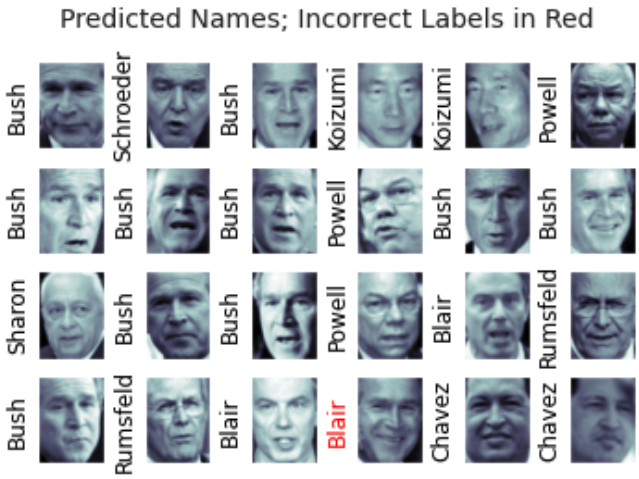


Figure 5.1: Example of face classification using SVC

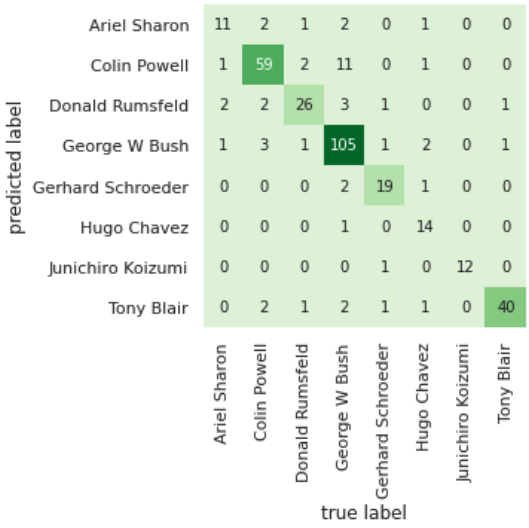


Figure 5.2: Heat map of face classification

	precision	recall	f1-score	support
Ariel Sharon	0.65	0.73	0.69	15
Colin Powell	0.80	0.87	0.83	68
Donald Rumsfeld	0.74	0.84	0.79	31
George W Bush	0.92	0.83	0.88	126
Gerhard Schroeder	0.86	0.83	0.84	23
Hugo Chavez	0.93	0.70	0.80	20
Junichiro Koizumi	0.92	1.00	0.96	12
Tony Blair	0.85	0.95	0.90	42
accuracy			0.85	337
macro avg	0.83	0.84	0.84	337
weighted avg	0.86	0.85	0.85	337

Figure 5.3: Accuracy of gender classification

5.2 Classification by Gender

We applied the SVC classification tool to the 512-dimensional embeddings generated from Google's Universal Sentence Encoder module.

Figures 5.4 and 5.5 are the classification heat map and accuracy metrics respectively. While accuracy measurements were not as high as that for face classification, there appeared to be something within the embeddings that was predictive of an individual's gender.

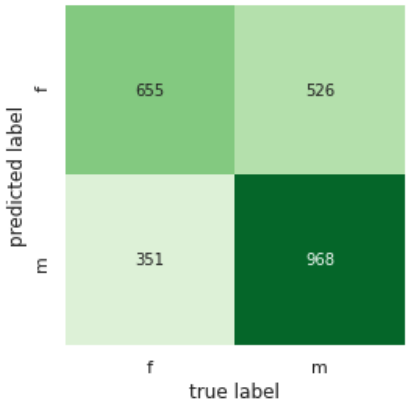


Figure 5.4: Heat map of gender classification

	precision	recall	f1-score	support
f	0.55	0.65	0.60	1006
m	0.73	0.65	0.69	1494
accuracy			0.65	2500
macro avg	0.64	0.65	0.64	2500
weighted avg	0.66	0.65	0.65	2500

Figure 5.5: Accuracy of gender classification

5.3 Classification by Age Group

Figures 5.6 and 5.7 show the classification heat map and accuracy metrics respectively. The classification of a person's age group was not very good in general.

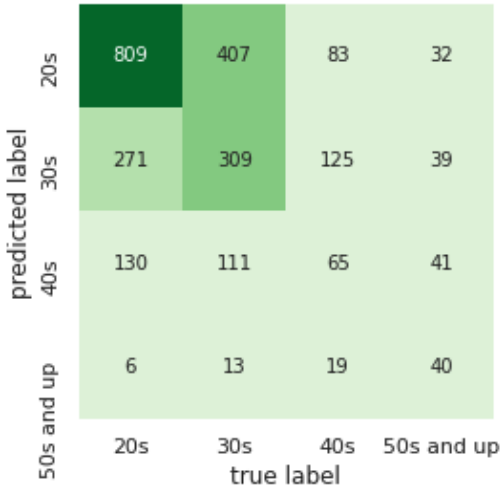


Figure 5.6: Heat map of gender classification

	precision	recall	f1-score	support
20s	0.61	0.67	0.64	1216
30s	0.42	0.37	0.39	840
40s	0.19	0.22	0.20	292
50s and up	0.51	0.26	0.35	152
accuracy			0.49	2500
macro avg	0.43	0.38	0.39	2500
weighted avg	0.49	0.49	0.48	2500

Figure 5.7: Accuracy of gender classification

CHAPTER 6

Correlation of Essays

6.1 By Age, by Gender, by Orientation

We can compare how closely related the word choices in dating essays are between different characteristics like men relative to women, younger daters versus older ones, high school graduates versus college graduates, etc. A perfect 1.0 correlation coefficient implies that the word frequencies are the same between two groups.

In Figure 6.1, the first bar (correlation coefficient = 0.97) indicates that men and women's dating essays compared very closely in their choices and frequencies of words used. It turns out that Venus and Mars are actually not that much different in the area of word frequency correlation.

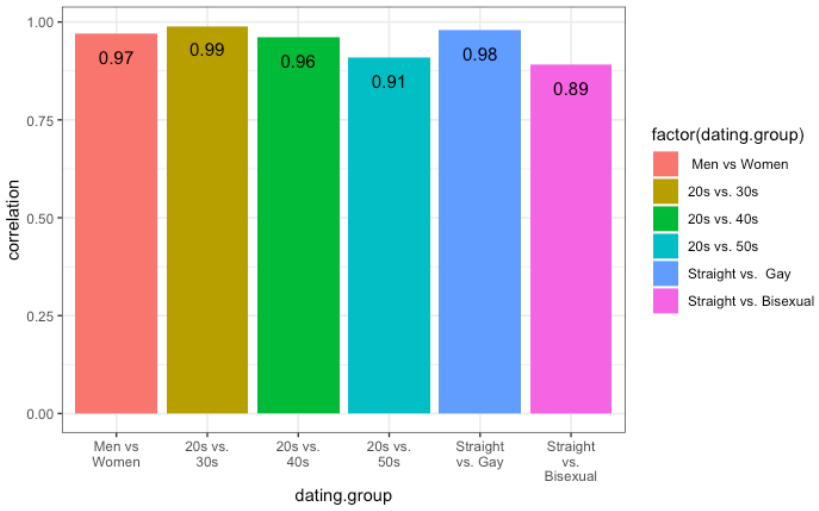


Figure 6.1: Essay correlations between age, gender, orientation

Turning now to how the younger folks compared to more mature individuals, we saw that the dating essays between the two groups are also extremely similar (correlation coefficient = 0.99) when comparing people in their twenties versus those in their thirties. This made intuitive sense since most people wrote about their interests and hobbies, and those tend to be quite similar when the age gap was only a decade apart on average.

As we increased the age gap to two decades apart i.e. comparing the self-summaries of the daters in their twenties versus those in their forties, the correlation—while still very high—dropped from 0.99 to 0.96 (third bar). With a 30-year age gap difference when comparing age group of twenties versus fifties, the correlation dropped another few points to 0.91 (fourth bar). This demonstrated the change in interests and hobbies from younger to older people and was interesting to observe the statistical evidence of this.

When it comes to sexual orientation, it appeared that the dating essays for the straight daters were very similar text-wise when compared to the gay daters: correlation of the written profiles between the two groups is at 0.98 (fifth bar). The comparison between straight daters versus bisexual daters showed a more noticeable difference, with the correlation calculated of 0.89 (sixth and last bar). The 0.89 correlation is also the lowest correlation measurement between exploration of dating profiles as a group versus another group.

6.2 By Education, by Racial Group

Were there any significant differences in dating essays when it came to educational levels completed, or when it came to comparing one ethnic group to another? It did not appear so, as shown in Figure 6.2. All of the different group comparisons are 0.97 correlation or higher.

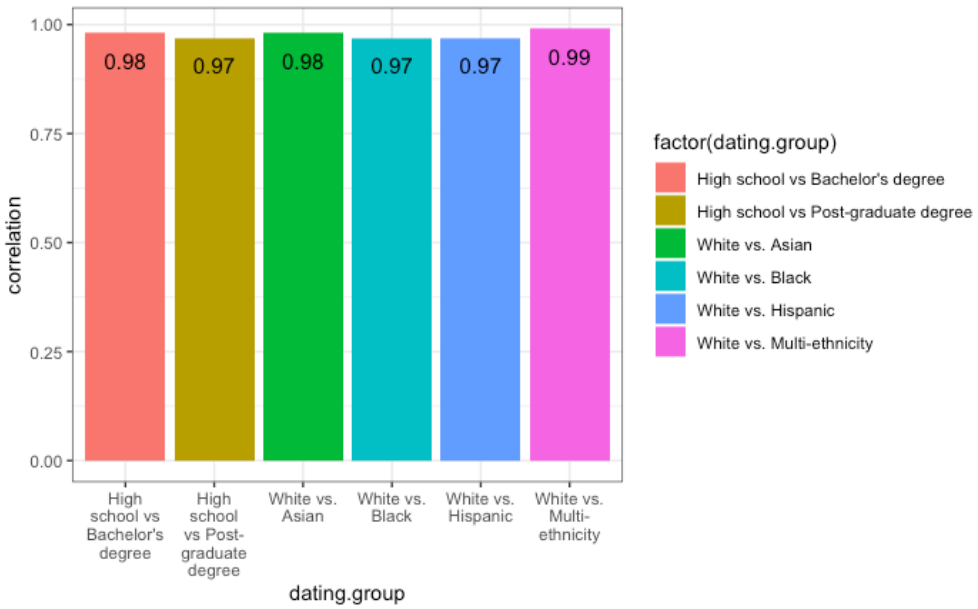


Figure 6.2: Essay correlations between education, racial groups

6.3 By Religion

Figure 6.3 summarizes the correlation coefficients of the dating essays between the major religious groups. The first bar comparing essay correlation between genders will serve as a good reference point.

Dating singles in this data set who have indicated their religious beliefs compared very closely in terms of their dating essays. All essay correlations between groups were 0.96 or higher.

Christian and Catholic daters were the most similar (correlation = 0.99), followed closely by Christian and Jewish daters (correlation = 0.98). The essays showed the largest difference between Christian and Atheist daters: correlation between these two groups was 0.92.

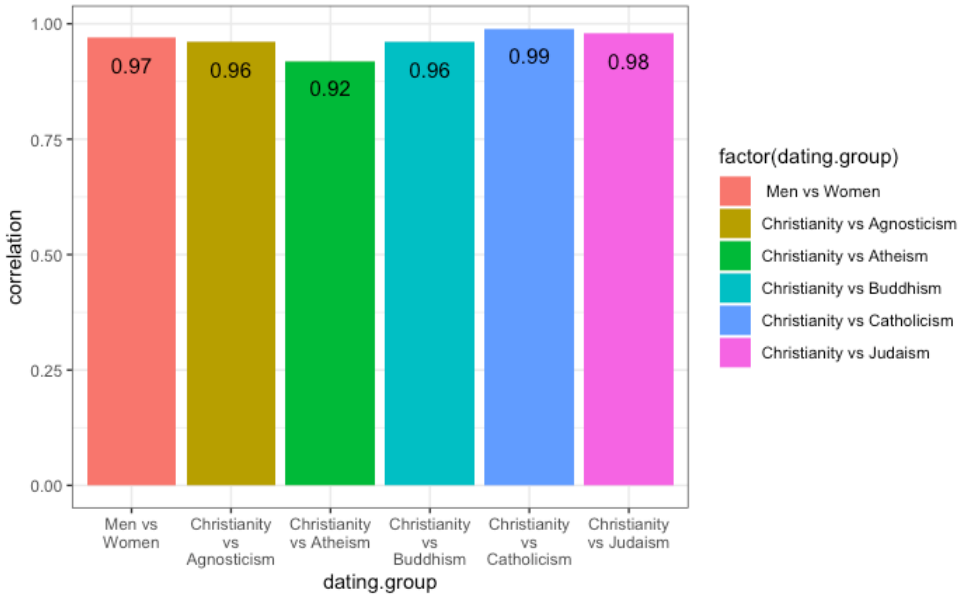


Figure 6.3: Essay correlations between religions

6.4 By Drink Habit, by Pet Preference

The essay comparisons for different drinking patterns (often, socially, or rarely/teetotaler) and diet preferences were very similar: the correlations compiled from this data set were 0.97 or higher.

Essays exhibited relatively larger word choice and frequency differences depending on which pet types they preferred. When we compared the essays of those who indicated that they like dogs versus those who like cats, the correlation dropped to 0.93. There appeared to have been more of a difference in word choices and/or word frequencies for cat fans versus dog fans.

It is worthwhile to emphasize that 0.93 remained a very highly positive correlation measurement. Any specific commentary here was for relative values.

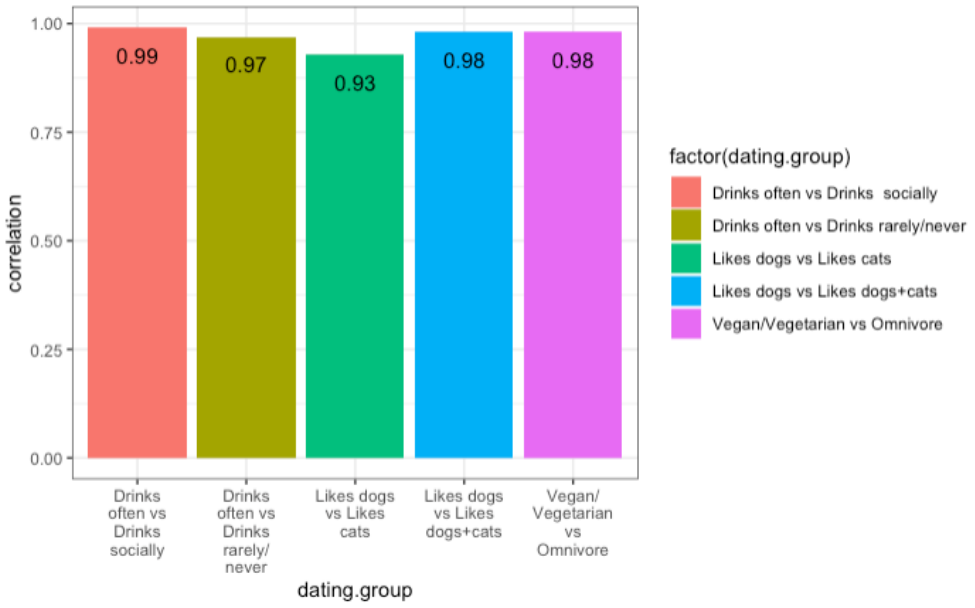


Figure 6.4: Essay correlations between drinking habits, pet preference

CHAPTER 7

Topic Modeling

7.1 Latent Dirichet Allocation

Topic modeling using latent Dirichet allocation did not reveal clear and meaningful topics (Figure 7.1).

This corresponded with the other statistical measurements evaluated so far: high correlation coefficients between essays of different groups of people, x-y plots of word choices and frequencies that fell very closer to the diagonal rather than axes. It appeared that there was a lot of overlap in the dating essays of the nearly 60,000 people in this data set.

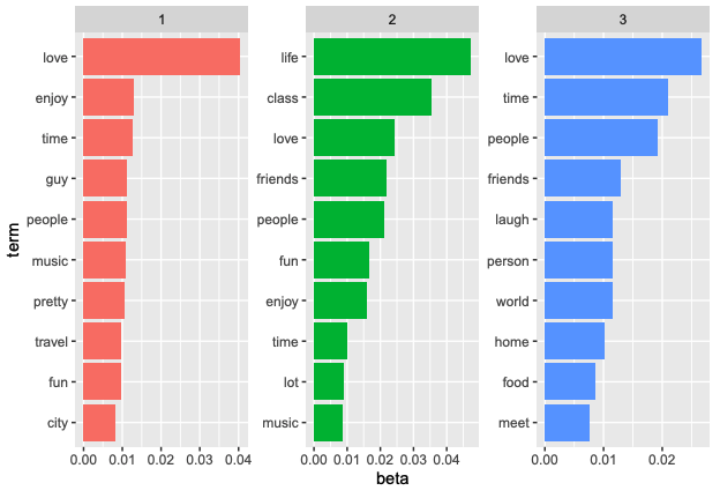


Figure 7.1: LDA topic modeling with k = 3

CHAPTER 8

Word Frequencies

8.1 Word Count by Age, by Gender

By counting the number of words in each self-summary and compare the essay length, there were slight differences by age and by gender as shown in Figure 8.1.

On average, the self-summary essay length of women (mean = 116) were a few words more than men (mean = 109). We can see this in the scatter plot of word counts (Figure 8.1) where the red fitted line representing women is slightly above the blue fitted line representing men.

The word count followed a small increase of a few words each year as the daters get older. The average word count for 20-year olds were about 80 while that for 40-year olds numbered around 120. This translates to an increase of roughly 20 words per decade of life.

It was indeterminate whether the word count increase was due to chronological age, or if generational differences explained the change. The data set was a snapshot from 2012, when the 20-year olds were born in 1992 while the 40-year olds were born in 1972. A longitudinal study that followed the 20-year olds as they aged into their 40s may help shed more light on the driver of the essay length increase with age.

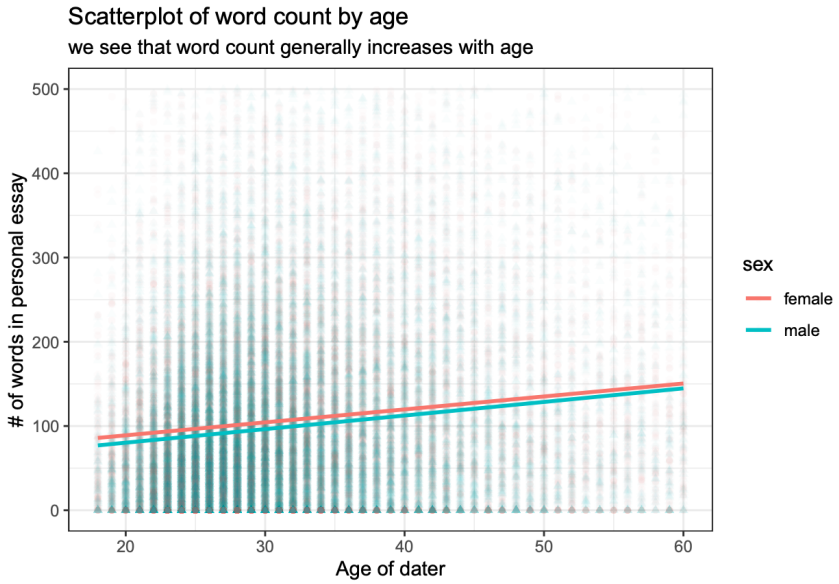


Figure 8.1: Length of dating profile essays by gender and age

8.2 Most Frequent Words

The most frequent word used was "love" as shown in Figure 8.2. This was hardly surprising, as daters were prompted to mention their favorite activities, food, books, movies, etc. to elicit responses from others with similar preferences.

The natural question that followed was to explore what words which were in close proximity in sentence order with "love". This makes use of a technique called "N-grams".

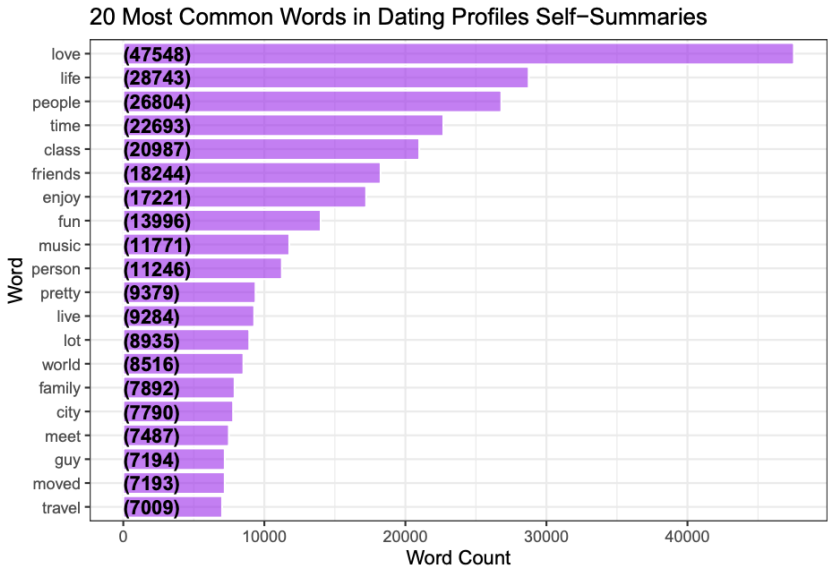


Figure 8.2: Most common words in dating essays

CHAPTER 9

Word Networks

9.1 Bi-grams

An n-gram is a segment of n consecutive words from a sample of text provided. A bi-gram is the shortest possible n-gram as it takes just two words.

By comparing the multitudes of bi-grams extracted from a text sample, we can compute relative frequencies of each bi-gram against the others and sort out which are the most frequently co-occurring word-pairs.

Bi-grams for "love" were interesting because the word is often used in conjunction with hobbies and favorite things in dating essays. Women appeared to enjoy food, nature, reading, cooking and movies (Figure 9.1). For men, their favorite activities seemed to relate to sports and music (Figure 9.2).

9.2 Word Networks

A few things of interest in the word networks for women self-summaries: the women appeared more interested in travel since "south america" was featured, and also perhaps more marriage-minded with terms like "family oriented" being mentioned (Figure 9.3).

On the other hand, men wrote more about their hobbies and activities: mountain biking, road trips, rock climbing and board games (Figure 9.4).

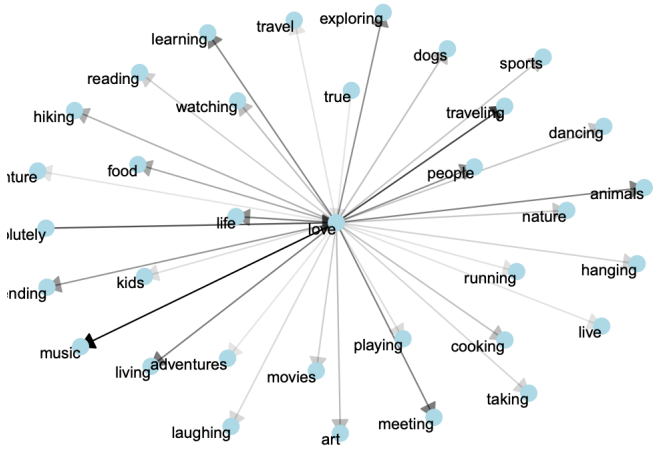


Figure 9.1: "Love" bi-grams in women's essays

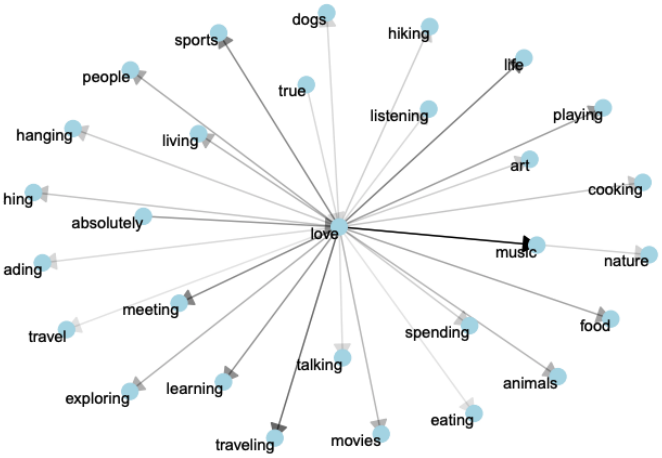


Figure 9.2: "Love" bi-grams in men's essays

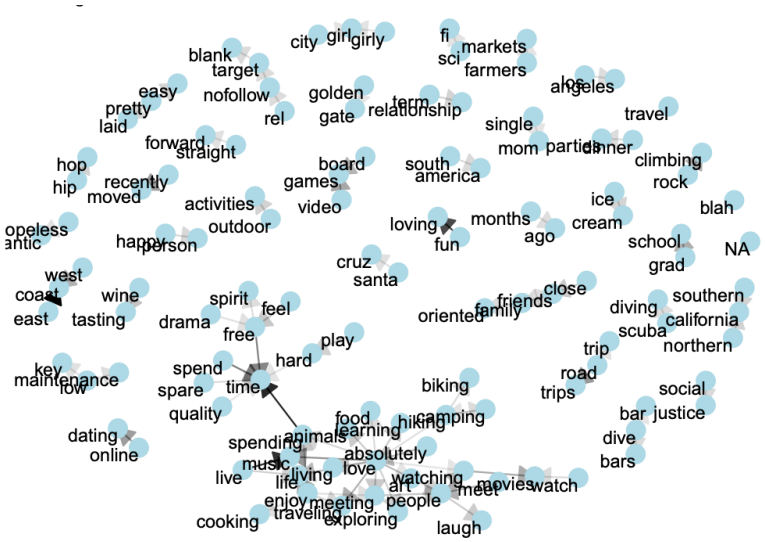


Figure 9.3: Word networks in women's essays

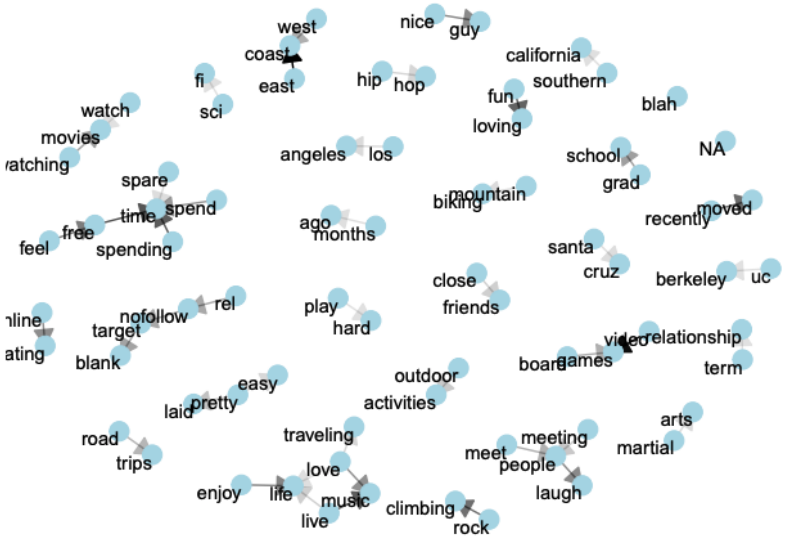


Figure 9.4: Word networks in men's essays

CHAPTER 10

Most Unique Words

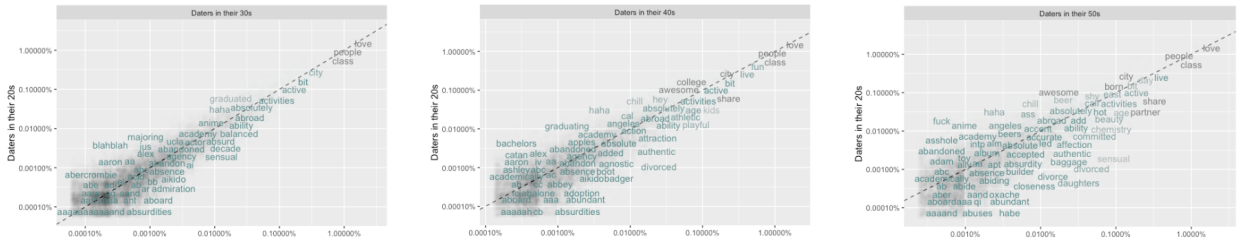
10.1 X-Y Plots of TF-IDF

What were the words that helped differentiate the men from women? Younger versus older daters? We reviewed this topic by transforming the essays into the term frequency-inverse document frequency or "tf-idf" format, then plotted the groups along different chart axes.

How would differences in word choices and frequencies look like in such a plot? If the dating essays for group A were exactly the same in both words used and their frequency, then all the words would lie exactly on the diagonal line on this type of x-y plot. If the essays share no common words, we will expect a plot of the words for group A to fall on its axis (say x-axis) and those for group B to fall on the other axis (y-axis in this example).

As an example, Figure 10.1 shows the word dispersion for three different age gaps: going from a 10-year gap, to a 20-year gap, and lastly to a 30-year gap. The visualization in a side-by-side format is clear: when the age gap is just 10 years the words of the two age groups remained relatively close to the diagonal.

As the age gap increased, we saw that the words started to peel away from the diagonal and fan out toward the axes. This is a visualization of the dating essays being more different between young and middle-aged people.



(a) 20s vs 30s (b) 20s vs 40s (c) 20s vs 50s

Figure 10.1: Side-by-side view of word dispersion as age gap increases

10.2 By Gender

From Figure 10.2, we see that the word choices and frequencies are very similarly when comparing men's dating essays with women's.

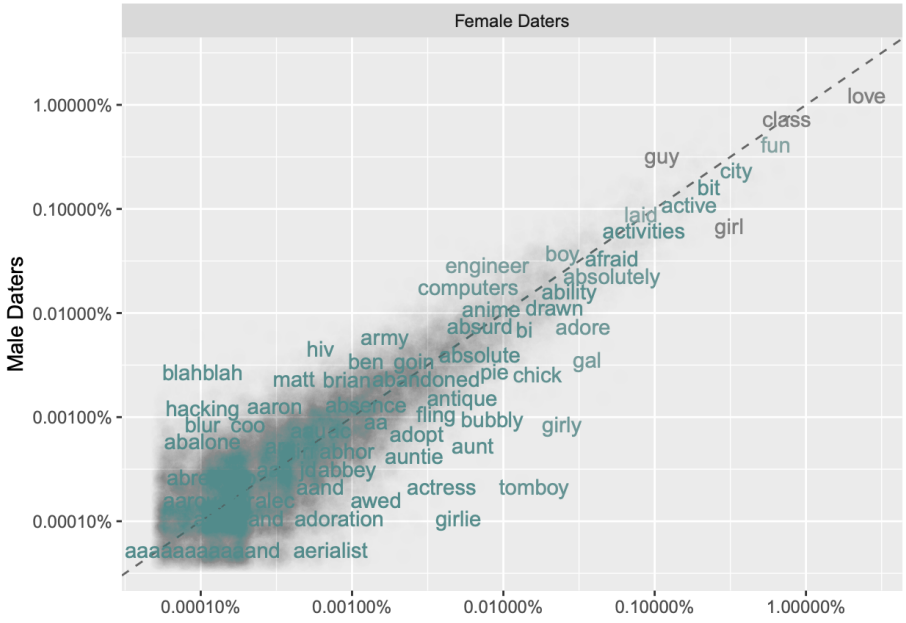


Figure 10.2: Compare word frequencies by gender

A few samples of the words that were distinctive to women in their self-summary essays were: actress, tomboy, girly, chick, adore, bubbly. Going over to the men, their more unique vernacular included: hacking, computers, engineer, lasers, outdoorsman, skateboarder.

The word list for men made intuitive sense since this data set is of nearly 60,000 daters in the San Francisco and surrounding area. The San Francisco area is well-known for being the technology hub of the U.S. and in the world, and the workers in the technology industry historically has been dominated by young Caucasian men, likely well-educated and at the higher end of the socio-economic spectrum.

The word "HIV" was mentioned in men self-essays more frequently than for women. Another relatively unusual characteristic about San Francisco is that a relatively large gay population reside there. The gay men likely wrote about their HIV-status—whether positive or negative—in their dating essays for responsible disclosure to their dating prospects.

10.3 By Age Group

Figure 10.3 is the x-y plot of word frequencies when comparing people in their twenties versus those in their thirties. The words were quite tightly clustered along the diagonal line, again indicating that the word choice frequencies were very similar between the two age groups. One of the words most unique to those in their twenties was "Abercrombie" which turned out to be a preppy clothing retailer favored by the youthful crowd in or just out of college.

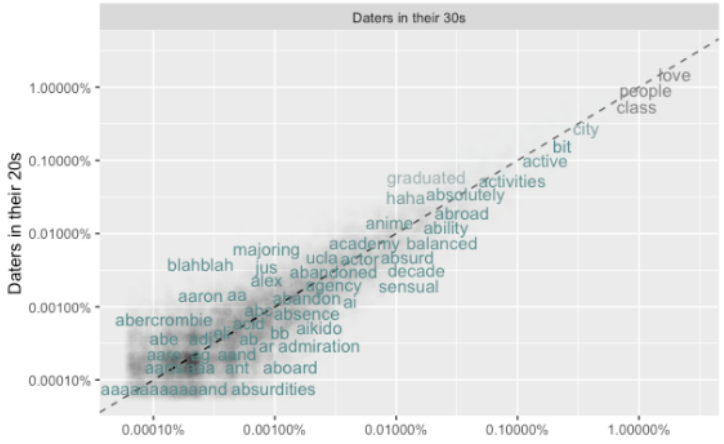


Figure 10.3: Compare word frequencies of age 20s vs 30s

Comparing the essays of daters in their twenties with those in their forties (Figure 10.4), we started to observe a shift in the words used and frequencies. For example, "divorced" popped up as a word more frequently used by individuals in their forties. The word "authentic" also came up as frequent and more unique for the over 40 crowd. This hints at an increasing maturity and seeking connections which are less about the superficial and more about deeper connections.

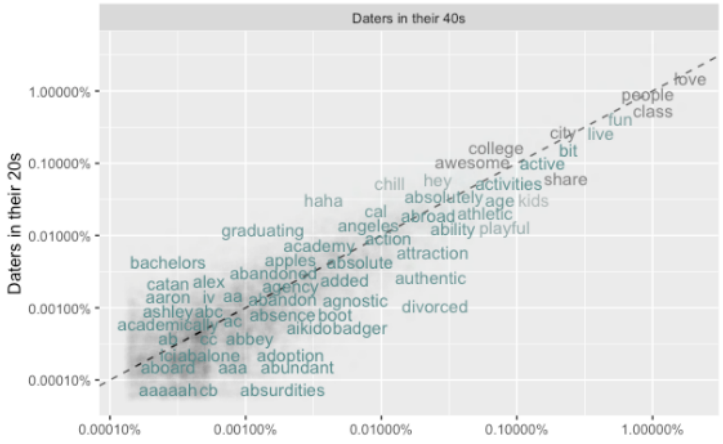


Figure 10.4: Compare word frequencies of age 20s vs 40s

When we increased the age gap to 30-year difference i.e. comparing word selection and frequencies between folks in their twenties versus fifties, a life-stage gap became evident in the plot (Figure 10.5).

The words used by each group and frequencies showed more dispersion away from the diagonal line. The vocabulary exhibited the even more differences with a 30-year age gap compared to when the gaps were smaller. The words below the dotted line are pertinent to the daters' in their fifties, and aligns with people who have started families (a word like "daughters") and likely gone through relationship break-ups (words like "divorced", "baggage"). They also seemed to have more spiritual aspirations for new relationships, using words like "committed", "share", "partner", "affection" in their dating essays.

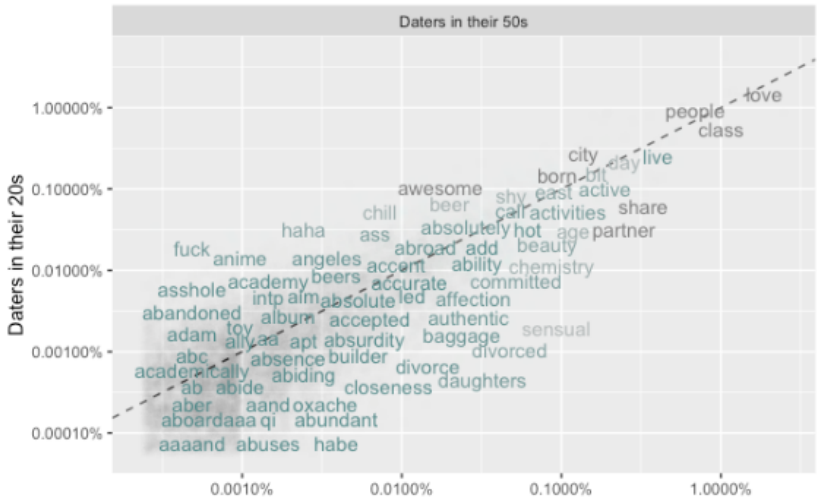


Figure 10.5: Compare word frequencies of age 20s vs 50s

10.4 By Pet Preference

We saw in essay correlation coefficients that the essays by dog-lovers were less correlated to those by those with a preference for cats. Figure 10.6 helps tease out the essay differences between the two pet-owner groups.

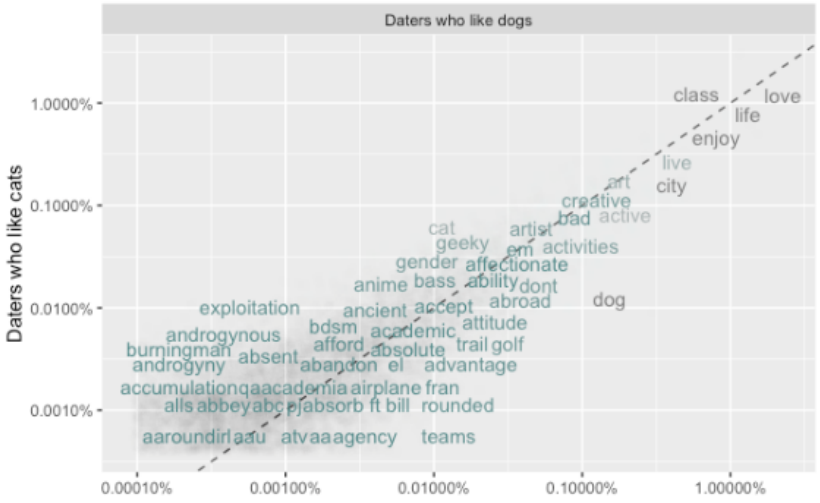


Figure 10.6: Compare word frequencies by pet preference

The word "dog"—positioned under the diagonal and closer to the x-axis—was one of the

words identified as being more unique and used more frequently in dog-lovers' dating essays. This was mirrored by the word "cat" which was above the diagonal and closer to the y-axis.

This came as no surprise and served as helpful validation that when essays were transformed, summarized and plotted this way, we can easily visualize whether essays are more similar or different.

Dog-lovers used words like active, activities, trail, golf, airplane, and abroad in their dating essays more frequently. This hinted at the dog-lovers possibly having more outdoorsy and upper socio-economic lifestyles.

On the other hand, cat-lovers seemed to have more alternative lifestyles. Words which were more unique to them included: burning man (a hippie festival), androgyny, BDSM (bondage, discipline, sadomasochism), and anime.

CHAPTER 11

Business Applications

We have gone over some insights found from text analysis. Are there ways that Tinder and other companies can apply this knowledge to the business of online dating? Here are a few possibilities:

- Essay length: There is likely an optimal essay length range analogous to the three little pigs children's story. Ideally essays should not be too short nor not too long. The dating apps can provide alerts when number of words in essay gets too too/high, and encourage their customers to be more editorial with their self-summaries.

- Sentiment analysis: The dating apps can provide alerts when sentiment analysis of their customers' self-summaries are too negative. After being alerted, the customers can choose to edit their profiles to emphasize their interests and hobbies.

- Most common words: Tinder, Match and OKCupid can consider organizing dating meet-up events centered around these words e.g. drinking, bars, wine, outdoors, travel, food, cooking, museum, art, etc.

- Most unique words: Highlight those and "coach" their customers, suggest good first message that ask about any unique words/hobbies. For example "I see you're a competitive birthday cake baker - that's so unusual - what did you bake in your last baking competition?"

- Essay correlations: Rank the essays by correlation measurements for each customer, and use this as a sort order in the recommendation engines.

CHAPTER 12

Conclusion

Text mining work on dating essays is fascinating from an anthropological and cultural perspectives. However, it is not without practical potential as discussed.

An exciting new area is the development of language generation bots. OpenAI, a San Francisco-based artificial intelligence research laboratory, is at the forefront of such development. OpenAI introduced the Generative Pre-trained Transformer 2 (GPT-2) in 2019, followed by GPT-3 in mid-2020.

The GPT models can take a text data corpus (books, news articles, customer reviews), apply deep learning, and then generate human-like text at a prompt. Imagine the business value of an essay-bot trained on the dating essay corpus of Tinder, Match, or Bumble.

The dating app user can start by contributing a few sentences about themselves. These sentences will serve as the starting prompt for the GPT model essay bot, which generates new continuation sentences. The Tinder user can choose whether to accept, edit or reject such essay-bot synthesized essay continuations.

While dating app users should all strive to present their authentic selves and write accurate, honest and erudite profiles, we know that not everyone is a wordsmith. A large proportion of single folks confess to finding it tough to talk about themselves. New tools like a GPT-2 essay bot can definitely help address these all-too-common hurdles for single people. This is a win-win proposition for the dating app industry and its users.

REFERENCES

Argamon, Shlomo et al. (2005). "Lexical predictors of personality type". In: *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*, pp. 1–16.

Buss, David M (1985). "Human mate selection: Opposites are sometimes said to attract, but in fact we are likely to marry someone who is similar to us in almost every variable". In: *American scientist* 73.1, pp. 47–51.

Hirsh, Jacob B and Jordan B Peterson (2009). "Personality and language use in self-narratives". In: *Journal of research in personality* 43.3, pp. 524–527.

Hu, Minqing and Bing Liu (2004). "Mining and summarizing customer reviews". In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177.

Kim, Albert Y and Adriana Escobedo-Land (2015). "OkCupid data for introductory statistics and data science courses". In: *Journal of Statistics Education* 23.2.

Mohammad, Saif M and Peter D Turney (2013). "Crowdsourcing a word–emotion association lexicon". In: *Computational intelligence* 29.3, pp. 436–465.

Nielsen, Finn Årup (2011). "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs". In: *arXiv preprint arXiv:1103.2903*.

Pew, Research Center (2020). *10 facts about Americans and online dating*. URL: `https://www.pewresearch.org/fact-tank/2020/02/06/10-facts-about-americans-and-online-dating/`.

Tausczik, Yla R and James W Pennebaker (2010). "The psychological meaning of words: LIWC and computerized text analysis methods". In: *Journal of language and social psychology* 29.1, pp. 24–54.

Yarkoni, Tal (2010). "Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers". In: *Journal of research in personality* 44.3, pp. 363–373.

Youyou, Wu et al. (2017). "Birds of a feather do flock together: Behavior-based personality-assessment method reveals personality similarity among couples and friends". In: *Psychological science* 28.3, pp. 276–284.