THE UNIVERSITY OF CHICAGO

What people worry about:

Top Personal Finance Concerns with Reddit Data[1]

By

Jinfei Zhu

June 2022

A paper submitted in partial fulfillment of the requirements for the Master of Arts degree in the Master of Arts in Computational Social Science

Faculty Advisor: Jon Clindaniel

Preceptor: Shilin Jia

[1] The data and codes for this paper can be found here: https://github.com/jinfei1125/personal-finance

Abstract

Personal finance problems have become a significant social issue that bothers Americans despite their high income. Financial concerns could lead to mental depression, anxiety, sadness, and lower productivity. To find out the reasons behind this phenomenon and discover the top personal finance concerns, this paper utilizes the posts from the online Reddit discussion forum *Personal Finance,* with TFIDF to find out the top words in 13 categories of the forum, LDA Topic Modeling to reveal the distribution and content of topics in the online discussion, and LIWC dictionary to detect the psychological traits under each flair.

Keywords: Personal Finance, Reddit, TFIDF, LDA Topic Modeling, Computational Social Science, Natural Language Processing, Sentiment Analysis, LIWC Dictionary

1. Introduction

In 2017, a Federal Reserve survey (Federal Reserve, 2018) finds almost 40% of American adults wouldn't be able to cover a $400 unexpected emergency expense with cash, savings, or a credit card charge that they could quickly pay off, saying that they would either not be able to cover it or would cover it by selling something or borrowing money. According to the Consumer Expenditure Surveys (CE) of the U.S. Bureau of Labor Statistics, the average monthly expenditures per household[2] is $5,005 in 2017, with $1,657 in housing (including both owned and rented) and $644 in food (Bureau of Labor Statistics, 2017), which indicates $400 is not a very large amount. At the same time, the United States ranks 5 in terms of GDP per capita (IMF, 2021) all over the world. Why are people in the United States, the country with one of the highest incomes per capita, so unprepared for financial emergencies?

The lack of emergency money and financial concerns could cause mental burdens such as worry, stress, anxiety, or sadness. According to an experiment by Kaur et al (Kaur et al. 2021), financial concerns would make workers less productive while on cash-rich days, employees work faster and make fewer errors. In a large-size random control trial in Bangladesh from more than 20,000 households, the result shows that there exist barriers for the poor, who are mostly involved in the low-skilled and seasonal jobs, to take the same work activities, livestock rearing, as the non-poor, even when they were provided with

[2] The average household size is 2.5 persons, which includes other age groups such as kids along with adults.

the same resources (Bandiera et al. 2017). Lehman and Koerner find that the financial hardship of mothers is significantly related to daughters' psychological depression (Lehman and Koerner 2002). Financial scarcity will also lead to less child-directed speech from parents, with a negative impact on children's later life development (Ellwood-Lowe, Foushee, and Srinivasan 2022).

For students, while facing expensive tuition fees and high living costs, their overall knowledge of personal finance was limited, which in turn led to doubts and anxiety for students (Mazhari and Atherton 2021). Another study conducted in Kenya suggests a causal effect of negative economic shocks on stress levels, by comparing both levels of the stress hormone cortisol and self-reported stress level of farmers and non-farmers with the absence of rainfall, which could be a random economic shock to farmers (Chemin, de Laat, and Haushofer 2013).

Burdens of financial stress can lead to poorer decisions, especially for low-income individuals. Development Economists find that financial pressures make people less patient, showing more impatient behaviors including spending more on short-term temptation goods such as alcohol and entertainment, taking high-interest loans repeatedly, and missing high-return investment opportunities, which may contribute to self-reinforcing poverty (Bartos et al. 2018). A survey also shows that people tend to make more present-biased intertemporal decisions before their payday (Carvalho, Meier, and Wang 2016). Though in some cases, financial constraints can lead to more rational decision-making to avoid paying a high surplus for goods, as shown by

an experiment of over 3,000 farmers in Zambia (Fehr, Fink, and Jack 2022). Evidence from a shopping mall in New Jersey, US, and farmers in Tamil Nadu, India shows that a projected financial expense could negatively affect the performance on other decision-making and reasoning tasks of lower-income individuals, who would suffer more from cognitive pressure. This negative effect doesn't show for higher-income individuals (Mani et al. 2013).

Under this condition, knowledge, and skills in personal finance management, budgeting, and investing are essential. Taft et al. find that financial literacy could lead to greater financial well-being and less financial concern (Taft, Hosein, and Mehrizi 2013).

Therefore, it's important to know and understand what people's financial concerns are so we can increase productivity, help individuals make more rational decisions, reduce mental depression, and improve financial well-being.

To find out the reasons that contribute to people's financial concerns, I come up with the following research question:

What are the most frequent personal finance concern topics?

In this paper, I will try to find out the most widely discussed personal finance concerns on the internet by analyzing related online texts, so we can know why 40% of American adults are so financially unprepared, in order to find solutions to overcome the challenges facing us.

In the past, to answer this kind of people-related question, we may have to turn

to the power of surveys to ask each interviewee to answer specific questions designed by researchers. These questions, though well-designed, could be limited by researchers' opinions and limit the participants' expression. But in the digital era, many people discuss their financial concerns online and post their thoughts, questions, and suggestions online. Reddit is an online forum where users can ask questions, comment on each other, and upvote posts and comments they like, on various sub-forums for different interests (Medvedev, Lambiotte, and Delvenne 2019). These subforums are called subreddits and Personal Finance Subreddit is one of them, with enormous active users and detailed posts, which are openly available for researchers to dig out the information behind the text data.

Unlike other social media data such as Twitter, and Facebook, on Reddit, you cannot follow a single user, but only follow a subreddit, showing the content-based feature for discussion (Medvedev, Lambiotte, and Delvenne 2019). The word limit for a post is 40,000 characters, which is roughly 8,000 English words, providing users with enough space to share their stories.

Social Media data can shed light on insights to study concerns and anxiety levels of users. A study on student-posted tweets shows the main concerns of engineering students are heavy study load, lack of social engagement, negative emotion, sleep problems, and diversity issues (Chen, Vorvoreanu, and Madhavan 2014).

While many anxiety-detection studies are conducted using Twitter data, the word limit can suppress users' expressions. Shen et al. utilize LIWC dictionary and LDA topic modeling to study the anxiety level on Reddit from four subreddit forums r/anxiety, r/healthanxiety, r/healthanxiety, and r/panicparty as anxiety group, with a control group of 25 other subreddits, showing significant difference in the topic words as well as n-grams between the anxiety groups and control group(Shen and Rudzicz 2017).

2. Conceptual Models

Most Important Topics and Words

To find out the most frequent personal finance topics, this paper makes the following assumptions:

1. The posts fully represent people's personal financial concerns. The most frequent personal finance topics in people's daily life will be discussed the most times in the subreddit.

2. If a topic is important, not only the number of posts about this topic will be large, but specific words about this topic will also show up more often in the corpus.

Based on these assumptions, we can first count the number of posts. Second, we can also count the number of words to find out the most important topic by tokenizing sentences and words. Tokenization is the process to turn long

paragraphs and sentences into a list of words.

However, simply counting the number of words could result in some problems. One of the most obvious problems is the length of posts is different, containing a different number of words. Therefore, we need to normalize posts. Second, if one word appears in all topics, it's hard to determine if it's significant to one specific topic.

Therefore, TFIDF (Term Frequency-Inverse Document Frequency) score is calculated to reflect the most important word in the corpus, which increases proportionally to the number of times that a word appears in the document and is inversely correlated to the number of documents in the corpus that contain the word (Rajaraman and Ullman 2011).

Though sharing the same idea, there are many slightly different formulas to calculate the TFIDF score: binary, raw count, term frequency, log normalization, double normalization, and double normalization with K. In this paper, I used the following measures:

$$tfidf(t, d, D) = tf(t, d) * idf(t, D)$$

$$tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$idf(t, D) = \log\left[\frac{1 + |D|}{1 + df(t)}\right] + 1$$

With:

- t: one term t (word t)

- d: one document d
- D: the set of all documents
- $f_{t,d}$: the frequency of term t in document d
- |D|: the number of documents in D
- df(t): the document frequency of term t, the number of documents that contain t

With TFIDF instead of the raw frequencies of occurrence of a token in a given document, we can scale down the impact of tokens that occur very frequently in a given corpus.

Topic Modeling

Topic Model is first proposed by Blei et al. in 2003 and now it has become the most widely used unsupervised machine learning model to analyze text data. It is a two-dimensional clustering model which holds two basic assumptions (Blei, Ng, and Jordan 2003; Blei and Lafferty 2006).

1. A document exhibits multiple topics, such as debt, employment, and housing.
2. A topic is a distribution of a fixed vocabulary. For example, the investing topic contains words such as ‘fund’, ‘invest’, and ‘stock’.

It would allow us to find different clusters or topics in online posts. To discover topics in different posts, we could apply the Topic Model to texts of subreddits.

The Latent Dirichlet Allocation (LDA) Topic Model is a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document.

LDA topic modeling will need to be pre-set with the number of topics. The number 13 is chosen as the number of topics because the Personal Finance Subreddit has 13 different flairs to tag posts such as Retirement and Investing.

By using some visualization tools, we could draw the distribution of each topic and see their distances to each other and their sizes.

Sentiment Analysis

Sentiment Analysis is a subdomain of natural language processing, studying people's emotions and attitudes toward a topic, an object, or an individual. It can be reckoned as a classification algorithm to label text data into two classes based on the data's sentiment polarity: positive or negative. It has three different levels in Sentiment Analysis: document level, sentence level, and aspect level. At the document level, the algorithm analyzes the document as a whole to classify the polarity of its main opinion, positive or negative (Medhat, Hassan, and Korashy 2014).

LIWC Dictionary

By choosing the words to use, people give information about their beliefs, fears, personalities, thinking patterns, and social relationships (Stone, Dunphy, and Smith 1966; Gottschalk and Gleser 1979). Words are linked to individuals' psychological emotions and health when they are written down (J. W. Pennebaker 1993).

LIWC (Linguistic Inquiry and Word Count) is a method for computerized text analysis based on word frequency in different psychological categories. Created in 1993 for the first version, it is developed by Pennebaker Conglomerates, Inc. with the latest version in 2015 (J. W. Pennebaker 1993; J. Pennebaker, Francis, and Booth 1999; J. W. Pennebaker et al. 2015).

The core of this software is LIWC Dictionary, which is composed of about 6,400 words, word stems, and select emoticons. Each dictionary entry additionally defines one or more word categories or sub-dictionaries. Empirical studies have shown that LIWC can successfully detect psychological meaning in different contexts, such as emotionality, and social relationships (Tausczik and Pennebaker 2010).

Compared to traditional sentimental analysis, which usually gives three labels, positive, negative, or neutral. LIWC dictionary can capture more detailed emotions and sentiments. For instance, there are five categories for the word cried: sadness, negative emotion, overall effect, verbs, and past focus. So when

the word cried shows up in the corpus, the scores of all of these five categories will be increased, instead of just being marked as negative in the traditional way.

In this paper, the LIWC 2007 version (Pennebaker et al., 2007) has been used.

3. Data and Methods

Data Source

Reddit is a social news platform, web content rating, and discussion website, recently including the live stream functions. On Reddit, there are many subreddits that are forums for a specific topic. People's online discussion is a good reflection of their real-life concerns and thinking.

This paper chooses Personal Finance (r/personal finance) to scrape texts. Created on February 9, 2009, right after the 2008 financial crisis, it now has more than fourteen million members and usually has around ten thousand members online. It’s a very active and large subreddit compared to other subreddits and well-organized. Every hour there are many new discussions. People post their concerns, seek advice, or share personal experiences. There is a very detailed wiki for this subreddit (Reddit 2021) listing a summary of suggestions for different ages of people.

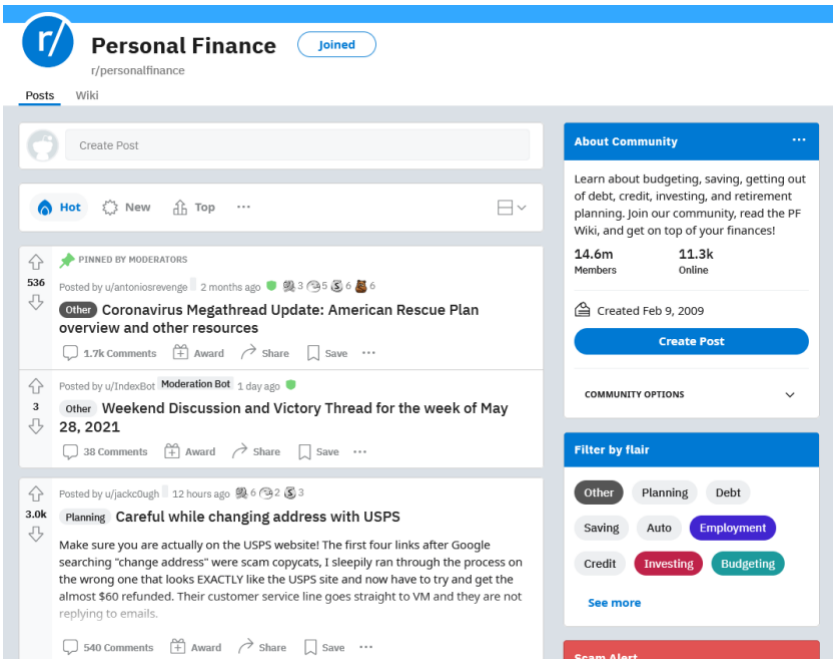

Figure 1 Home page of Personal Finance Subreddit

By browsing some posts, readers can easily find that there are a large group of anxious people—hot topics include paying back student loans, buying houses or cars with loans, taking care of aging parents, worrying about retirement, etc. These users ask for advice, offer their kind suggestions, and exchange ideas, helping each other solve financial concerns. Therefore, this subreddit would be a good choice to answer my research question.

Reddit has an API that allows developers to scrape posts and comments by different categories, such as ‘hot’, ‘new’, and ‘top’. But Reddit API has a limitation of no more than 1000 posts scraping each time, which is far more than a normal person’s reading ability. Therefore, I turned to Pushshift Reddit Dataset (Baumgartner et al. 2020) which offers Reddit comment and submissions archives.

The number of posts and users from 2009 to 2020 are presented in figure 1 and it shows that there is an S-shape for the growth rate. It starts to increase rapidly around 2013-2014, probably due to the popularity of the internet and smartphones, and then the growth becomes steady around 2017. In 2020, the COVID-19 pandemic caused a lot of shutdowns and unemployment, and the number of posts increase again.

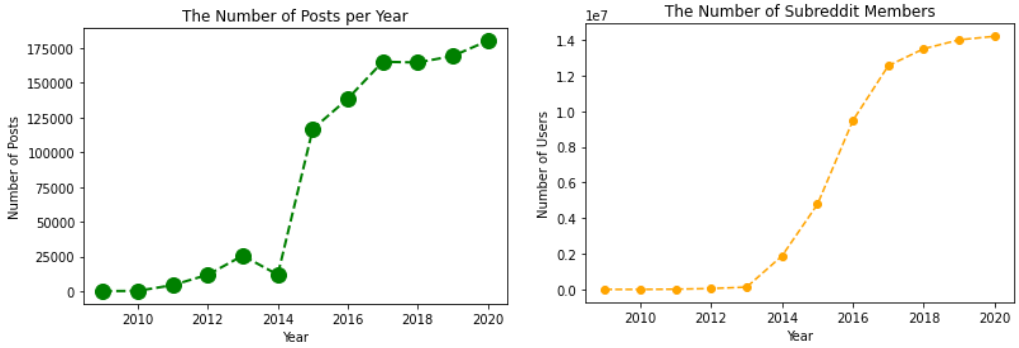

Figure 2 The number of posts and members of the Personal Finance Forum

The number of users also increases greatly from only 22 in 2009 to over 14 million in 2020, with a similar growth rate according to the increase of posts per year.

This S-curve growth reflects the result of user migration across different social media platforms. Since users are helping those websites' growth by posting, replying, and sharing, activities that produce valuable information on the platform, websites have incentives to attract users from other platforms. In 2011, Kumar studies the migration among Delicious, Digg, Flickr, Reddit, StumbleUpon, Twitter, and YouTube (Kumar, Zafarani, and Liu 2011) with data from March to May 2010 on these platforms. The result shows Reddit has a

small size of the potential migration population around 2010. Global social media research shows that during 2015, there were 2.3 billion active social media users in total, from which 2 billion users were active mobile users (Chaffey 2016). In 2016, Reddit released its mobile application and its download number has been steadily increasing (Reddit 2022).

In Personal Finance Subreddit, people discuss various personal finance topics, such as paying back debt, managing their credit scores, buying cars, comparing insurance, making investments, and so on. To make it easy for users to find posts that they are interested in, there is a 'Filter by Flair' box on the subreddit's home page, which contains 13 categories. These categories are shown in Table 1. This information is provided by the Reddit API. Since each post can only be tagged with one flair, we can first count the number of posts under each category to find the most popular topic, and then we can use these categories as a reference to find out the top words in each flair and decide the number of topics of posts in the topic modeling.

Table 1 'Filter by Flair' categories

Auto	Investing	Budgeting	Planning	Credit	Housing	Insurance
Retirement	Debt	Saving	Employment	Taxes	Other	

Comments constitute an important part of the discussion on Reddit. In some subreddits, the comments are even more important than the posts. However,

given the special feature of the Personal Finance forum, posts are usually long enough and contain key information—what people's personal finance concerns are. Therefore, my data only contain the "selftext", which is the original post apart from comments, on each Reddit post. These original posts, compared to their answers, are more likely to actually contain the concerns and detailed descriptions, whereas comments will generally contain "answers" to these questions. For example, here is a sample original post in 2020:

> I'm interested in the laws around eliminating PMI on my mortgage.
> My original mortgage was for 140,000. My property appraised out at 169,000, but I paid 155,000 (I negotiated hard!). I currently owe 126,000.
> My automatic PMI removal date was set for Feb 2024 based on just paying the minimum payment -- however, I have been paying aggressively.
> From what I understand, the bank is obligated to stop charging for PMI once I reach 78% of my homes original appraised value. I reached 78% over a year ago. I spoke to someone at USbank and they said it doesn't work that way; that the original cost of the purchase is considered and not the appraised value, and that I have to wait until 2024, pay down more (?) or have another appraisal done. Honestly, I don't totally understand what they meant.
> Does any know what the issue might be? Is it not as straightforward as the internet might have me believe? :)

Sampling Texts

The average length of posts is about 800 characters and 150 words. Since we have enough posts (989,193 posts in total), I use the systematic sampling method to sample 10% of the original data each year for my analysis (98,919

posts in total). The logic behind this sampling is to speed up the process of cleaning the data and building the model while maintaining the original ratio of the data year-over-year.

Data Cleaning

First, I dropped data rows that don't contain enough information, such as posts marked as 'removed', with None values, and with 'unset' flair.

Second, I removed stopwords and punctuation. Stopwords are words that don't have enough self-meaning but appear many times in the corpus, such as 'and', 'the', and 'I'. If we forget to remove stopwords from the corpus, the top words would be all stop words. I use the vocabulary list of the NLTK python package which includes 179 words, and add some Personal Finance specific stop words such as fractions of web URL 'r/personalfinance', 'www', 'com' etc.

Lastly, I stemmed the words to reduce repetition. Stemming is the process to derive words to their root form, so there is no impact from grammar numbers and tenses. For example, 'year' and 'years' would be counted twice as 'year'.

4. Results

Word Frequency Analysis with TFIDF

First, we can get a basic understanding of the number of posts under different flair classes. Figure 3 shows the results of all posts. Among them, the top concern is debt, which has more than 13,000 posts. In this sample, six

categories have more than 10,000 posts, which are debt, retirement, housing, credit, investing, and taxes. The least discussed topic is ‘budgeting’, this topic is a general term and can be further divided into clearer categories such as ‘housing’, ‘auto’, and ‘saving’. Besides, ‘budgeting’ sometimes has a similar meaning to ‘planning’, which could lead to different tags.

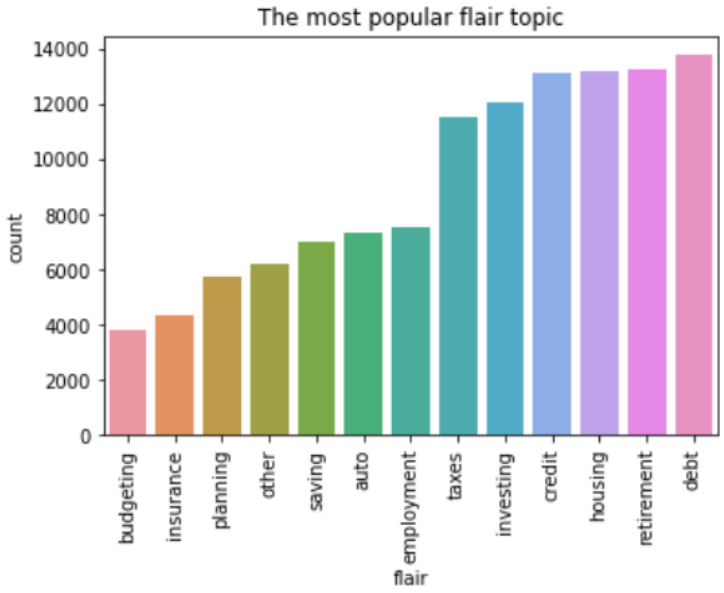

Figure 3 The most popular flair topic

When we look into top words in different flairs, there are many similarities across different flair classes.

Table 2 Top 5 words in 13 flairs [3]

auto	budgeting	credit	debt	employment	housing
car	save	credit	loan	job	house
pay	month	card	**pay**	work	year
year	year	**pay**	debt	year	home

[3] The full list of top words can be viewed on the GitHub repository.

loan	**pay**	score	credit	**pay**	**pay**
month	money	account	year	company	loan

insurance	**investing**	**other**	**planning**	**retirement**	**saving**	**taxes**
insurance	invest	account	year	ira	account	tax
year	fund	money	save	roth	save	year
pay	year	bank	money	contribute	money	file
plan	money	card	**pay**	401k	bank	**pay**
month	account	like	month	year	year	income

The most obvious outcome of this result is the word PAY appears almost in every flair—people are paying too many things: auto, budgeting, credit, debt, employment, housing, insurance, planning, and taxes—no wonder they don't have enough money for emergencies. Besides, other words that appear frequently across different flairs include YEAR, MONTH, and LOAN, which are also payment-related words.

Another interesting result is that for the "Retirement" flair, we can find there is a heated discussion about two different retirement plans: IRA and Roth 401K, where IRA is the individual retirement plan and Roth 401K is the retirement savings plan offered by employers.

Figure 4 Word Cloud of Personal Finance words

Figure 4 is a word cloud of all words across 13 flair classes. The top word is PAY, followed by time-related words such as YEAR and MONTH and TIME, which is in accordance with our findings in each flair. Other words also appear many times in the corpus, such as WORK, CAR, ACCOUNT, INCOME, and CREDIT CARD.

Topic Modeling

For topic modeling, I first used the Python Gensim package to create the LDA Topic Model based on the corpus, with 13 topics, then I used pyLDAvis, which is developed by Sievert and Shirley in 2014 to visualize the result of topic modeling, plotting the distribution of topics and words (Sievert and Shirley 2014).

The visualization is shown in figure 5 and 6:

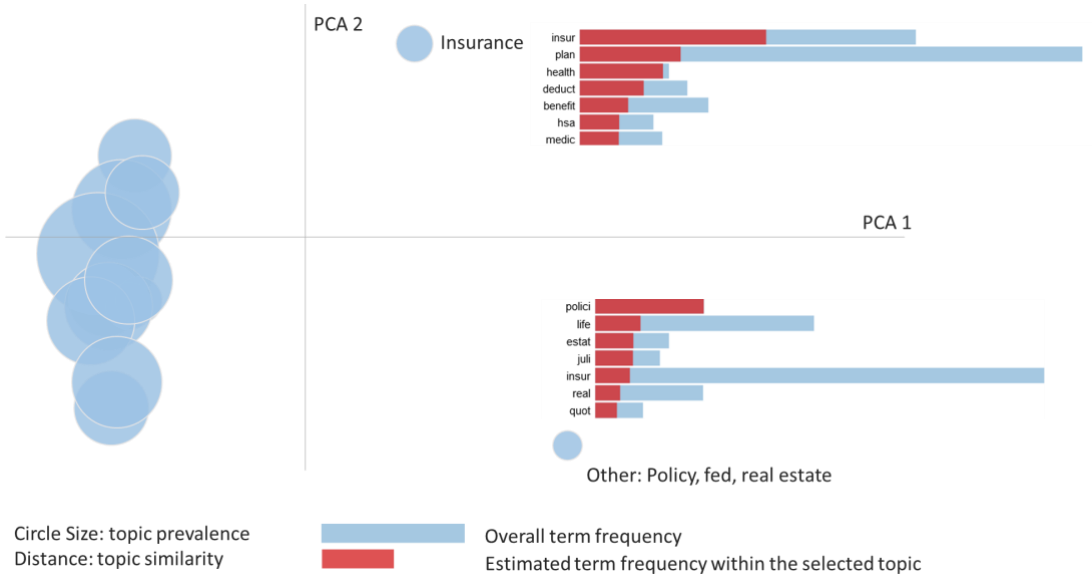

Figure 5 Digging out information behind two outliers

In this plot, each topic is presented as a circle. Principal Component Analysis is used, and the circles are plotted by the first two Principal Components.

The size of circles indicates the topic prevalence—the larger the circle is, the more prevalent the topic is in the corpus. The distance shows the similarity of topics. In the result, we find that 11 out of 13 topics are near each other—with almost the same value in Principal Component 1 and varies only in Principal Component 2.

By looking at the words in the two outliers which are far away from other topics, I found the topic in the upper right is about health insurance, and the one in the lower right is a mix of words and is hard to conclude a topic, such as POLICY, REAL ESTATE, QUOTE, and LIFE, so I labeled it as “other”. These two circles are not very big, so they are not very prevalent in people’s online discussions compared to other topics.

The blue bar and red bar for each word reflect the overall term frequency and estimated term frequency within the selected topic. For example, in the upper right health insurance topic, for the most frequent word INSURANCE, the length of the blue bar is almost twice the length of the red bar, which indicates that INSURANCE also shows up in other topics, because apart from health insurance, there are auto insurance and dwelling insurance. However, the word HEALTH almost exclusively exists in this topic.

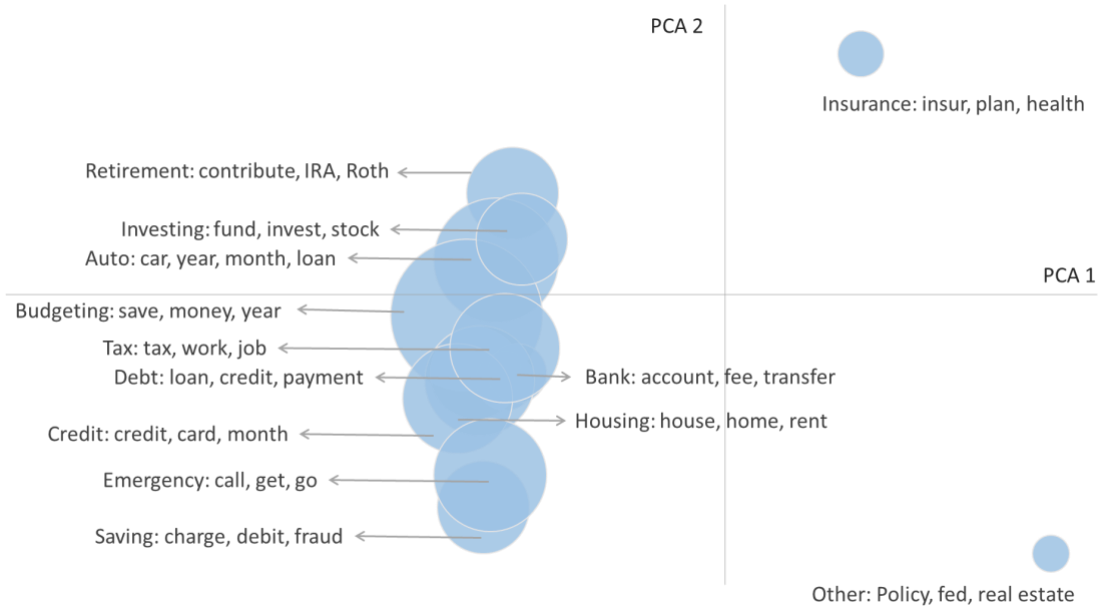

Figure 6 eleven topics are really near

The bulk of discussions, as figure 6 shows, are near each other. By looking into the top words in each cluster, I manually label their topics. They are Retirement, Investing, Auto, Budgeting, Tax, Debt, Bank, Housing, Credit, Emergency, and Saving. Among them, the group of Retirement, Investing and Auto have a positive value in Principal Component 2 and the rest have a negative value. Even though we could not assign the meaning of the principal components,

some possible sources of this similarity could be payment or saving related.

Besides this similarity, another thing to notice is that these topics have a large overlap of the original flairs, which means that the 13 manually set flairs could cover everything that users want to post. There are only two exceptions: the original flair system contains Employment and Planning, and in the results of topic modeling results, there are Emergency and Bank instead.

Sentiment Analysis

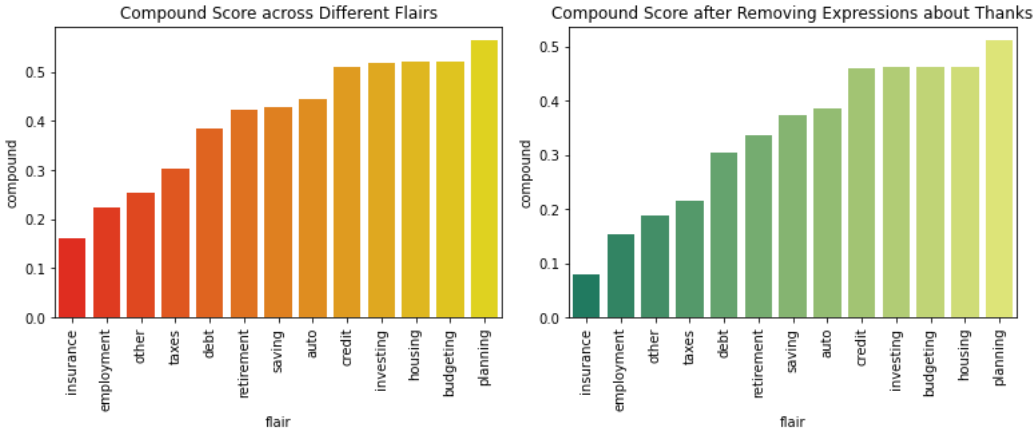

Figure 7 Compound Scores of Each Flair

We use basic natural language processing tools to calculate the compound score of each flair. After calculation, we would have three scores representing the positive, negative, and neutral sentiments in the original posts. By summing up these three scores as a compound score and normalizing it between -1 (purely negative) and 1 (purely positive), we can have a rough understanding of the sentiments of the text under each flair. Note that because of the friendly online discussion environment, each post contains many softening words like “Thanks” and “Appreciate”, which are not part of the main ideas but turn the

overall sentiment more positive. After removing thanks-related words, the overall sentiment scores are lowered by 0.05-0.1 for each flair, but the compound score for each flair is positive, indicating that users generally talk positively and friendly about their personal finance concerns. They are most negative when talking about Insurance and Employment, but they still have a compound score greater than 0. People are most positive when talking about topics on Planning, Housing, Budgeting, and Investing, because this usually means that they have extra money to do these financial activities, instead of worrying about which insurance is the best and the next job.

LIWC Dictionary

The results of the LIWC dictionary are similar to pure sentiment analysis, however, the result is slightly different. With LIWC, Credit is the most positive category and Tax is the least positive category.

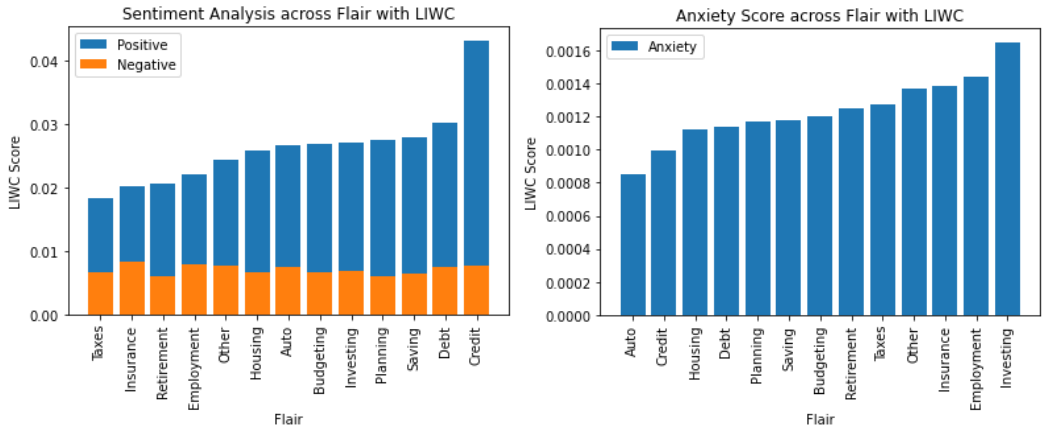

Figure 8 LIWC Sentiment Analysis and Detecting Anxiety

If we only look at certain psychological traits such as Anxiety, we find people are more anxious about Investing and Employment, in other words, the two

major sources of income – capital and labor. People are less anxious about Auto, Credit, and Housing, probably they are confident when they have money assigned to their cars and houses.

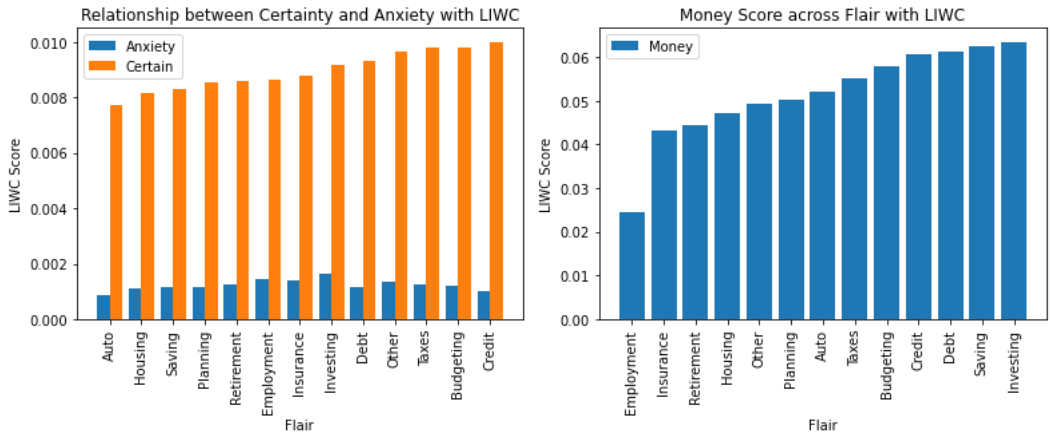

Figure 9 LIWC Result: Certainty, Anxiety, and Money

If we make a basic assumption of a potential negative relationship between Anxiety and Certainty since most people are risk-averse, the more certainty talked about in this flair, the fewer anxiety people may have. The result shows that this relationship doesn't exist in our current data.

LIWC also provides a possibility to count the frequency of money-related words. As the result of Figure 9 (right) shows, unsurprisingly, the top four flairs talking about money are Investing, Saving, Debt, and Credit, all more directedly related to money than other flairs.

5. Discussion

This paper takes advantage of the observational data from online discussion and natural language processing models to find out people's financial concerns,

which is a convenient and low-cost way compared to the traditional survey method. Besides the power of the computational method, there are three main strengths of this paper:

First, this paper is based on big data**,** with 98,919 posts (10% of 989,193 posts in total) from 2009 to 2020 and the average length of each post of 150 words, to analyze the top words in each flair and conducts topic modeling. This could reduce the randomness of small turbulence. Though big data will not help to decrease systematic error, instead, the impact of systematic error will become more obvious with the increase in data size. The most significant systematic error in my data could be the participants' age group. However, in my later analysis, I will prove this doesn't impede us from answering this paper's research question.

Second, Reddit posts have the property of always-on. People continue to post on Reddit and their posts will be kept on the website and archived for us to analyze, so we could collect the data within a large time range. In the future, I could also use this data to analyze the top words in each year and analyze the impact of a specific event by analyzing the posts before, in, and after the event, such as the COVID-19 pandemic.

Thirdly, observational data is less likely to suffer from bias due to the interaction of participants. This data is observational and it's less likely to change participants' behaviors. Besides, Personal finance concern is a relatively private

issue, so people rarely talk about it in their real life or feel awkward answering an anonymous survey, but thanks to the anonymity of internet forums, we can get posts about this important information. What's more, when people fill out survey questions, they may be constrained by the research questions and could not freely express their opinions. Reddit provides a natural setting that allows participants to tell their stories honestly.

In the meantime, this paper also suffers several possible biases. First, there probably is a generalization bias in my data. Most users of online platforms are young people who are used to the internet. Middle-aged people may not be willing to disclose their financial concerns online due to their unfamiliarity with online forums. However, a lot of young people are describing their parents' financial problems and ask for advice to help them. So, although there may not be enough direct information about the financial concerns of middle-aged or old people, there are many posts that indirectly mention them.

Since the data is not generated for research purposes at first, it may also suffer from incomplete information problems. Despite the abundance of texts, the data miss a lot of information for research. For example, there are no demographics in the data, so we don't know participants' age, gender, job title, industry, annual income, location, education level, marital status, etc. We don't have information about what contributes to these personal financial concerns. This is the disadvantage of observational data. A survey conducted by Schor showed that credit card debt has become one of the main reasons for financial stress—

people tend to underestimate their credit card debt by about 50% and many of them don't know where they spend their money (Schor, J.B. 1998; Schor 2008). However, we can't get this information from online texts, since the generation of these texts are not for research purpose, though we could accompany some survey results to help us analyze our data.

6. Conclusion

To answer the research question of this paper—what the most frequent personal finance concern topics are, I used two models to analyze the scraped corpus from the Personal Finance Subreddit: Counting words in each flair with TFIDF scores and LDA Topic Modeling.

The most significant finding of counting words is that people are overwhelmed by debt: the word PAY appears in the top 5 words in almost every flair. People need to pay a lot of things: credit cards, auto loans, housing loans, student loans, insurance, retirement, income tax…

Economist Juliet Schor mentioned in her book *The Overspent American* mentions that people are spending more than they did in the past and more than they realize. The saving rate has been around 8% for almost 10 years (the saving rate was even lower at the start of the 21st century)[4], which was lower

[4] According to the U.S. Bureau of Economic Analysis, the saving rate reached an almost historically high of 33.7% in April 2020 due to the shock of the pandemic, though it has fallen to 14.9% in April 2021.

than many countries of comparable income levels. Many families are living without an adequate financial cushion. 60% of all families' financial assets (savings outside of houses and cars) could last them about only one month if they lose their jobs (Schor, J.B. 1998). During recent COVID-19 pandemics, we have witnessed a great number of people lose their jobs and had to rely on unemployment benefits for a living.

Secondly, retirement has received a lot of discussions online. People are comparing individual retirement plans and retirement savings plans provided by employers.

Thirdly, the results of LDA topic modeling show that many personal finance topics, though they are discussed in different aspects, share a lot of similarities. While the distribution of Health Insurance locates far from other topics, other topics including Retirement, Investing, Auto, Budgeting, Tax, Debt, Bank, Housing, Credit, Emergency, and Saving are really near each other. This could give us a hint about Personal Finance Education—we don't need too many different strategies for different personal finance problems, since these problems are similar.

The result of LDA topic modeling has a large overlap of the categories of flair, which is set by the subreddit moderators. There could be two possible explanations for this phenomenon: first, the forum moderators are so good at summarizing the topics based on historical posts, so they set the flair that forum

members could always find a suitable flair to tag their posts; second, when users post their thoughts and they notice that they need to choose a flair to tag their posts, they unconsciously limit the content of their posts to the tag. The first explanation could make more sense, which is that the forum has a good flair tagging system and users can always find a suitable flair. Because when users write a post, choosing a flair is the last thing before they click the button 'Post'. In fact, it is easy to neglect the dropping box of choosing a flair until users received the warning that they cannot post without choosing a flair.

While the result of sentiment analysis shows that people generally hold a positive attitude toward their personal finance problems, for all topics, no matter it's about planning the beautiful future or some heavy topics such as unemployment. However, LIWC dictionary analysis shows that people are most anxious about two flairs, Investing and Employment, two sources of income. So if we want to mitigate people's personal finance concerns, more attention should be paid to increasing labor and capital returns.

Researcher Robert J. Gordon in his book *The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War* warns the younger generation may be the first in American history that fails to exceed their parents' standard of living(Gordon, 2016). The heavy personal financial burden is definitely one reason leading to this. This paper contributes to understanding the prevalent personal finance topics on Subreddit Personal Finance and could hopefully provide suggestions for personal finance educators and policymakers.

References

Bandiera, Oriana, Robin Burgess, Narayan Das, Selim Gulesci, Imran Rasul, and Munshi Sulaiman. 2017. “Labor Markets and Poverty in Village Economies*.” *The Quarterly Journal of Economics* 132 (2): 811–70. https://doi.org/10.1093/qje/qjx003.

Bartos, Vojtech, Michal Bauer, Julie Chytilová, and Ian Levely. 2018. “Effects of Poverty on Impatience: Preferences or Inattention?” SSRN Scholarly Paper 3247690. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.3247690.

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. “The Pushshift Reddit Dataset.” *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May): 830–39.

Blei, David M., and John D. Lafferty. 2006. “Dynamic Topic Models.” In *Proceedings of the 23rd International Conference on Machine Learning*, 113–20. ICML ’06. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1143844.1143859.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Carvalho, Leandro S., Stephan Meier, and Stephanie W. Wang. 2016. “Poverty and Economic Decision-Making: Evidence from Changes in Financial Resources at Payday.” *American Economic Review* 106 (2): 260–84. https://doi.org/10.1257/aer.20140481.

Chaffey, Dave. 2016. “Global Social Media Research Summary 2016.” *Smart Insights: Social Media Marketing*.

Chemin, Matthieu, Joost de Laat, and Johannes Haushofer. 2013. “Negative Rainfall Shocks Increase Levels of the Stress Hormone Cortisol Among Poor Farmers in Kenya.” SSRN Scholarly Paper 2294171. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.2294171.

Chen, Xin, Mihaela Vorvoreanu, and Krishna Madhavan. 2014. “Mining Social Media Data for Understanding Students’ Learning Experiences.” *IEEE Transactions on Learning Technologies* 7 (3): 246–59. https://doi.org/10.1109/TLT.2013.2296520.

Ellwood-Lowe, Monica E., Ruthe Foushee, and Mahesh Srinivasan. 2022. “What Causes the Word Gap? Financial Concerns May Systematically Suppress Child-Directed Speech.” *Developmental Science* 25 (1): e13151. https://doi.org/10.1111/desc.13151.

Federal Reserve. (2018). Report on the Economic Well-Being of U.S. Households in 2017.

Fehr, Dietmar, Günther Fink, and B. Kelsey Jack. 2022. “Poor and Rational: Decision-Making under Scarcity.” *Journal of Political Economy*, April, 720466. https://doi.org/10.1086/720466.

Gottschalk, Louis A., and Goldine C. Gleser. 1979. *The Measurement of Psychological States Through the Content Analysis of Verbal Behavior*. University of California Press.

International Monetary Fund (2021). World Economic Outlook Database. April 2021.

Kumar, Shamanth, Reza Zafarani, and Huan Liu. 2011. “Understanding User Migration Patterns in Social Media.” In *Twenty-Fifth AAAI Conference on Artificial Intelligence*. https://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/view/3664.

Lehman, Stephanie Jacobs, and Susan Silverberg Koerner. 2002. “Family Financial Hardship and Adolescent Girls’ Adjustment: The Role of Maternal Disclosure of Financial Concerns.” *Merrill-Palmer Quarterly* 48 (1): 1–24.

Mani, Anandi, Sendhil Mullainathan, Eldar Shafir, and Jiaying Zhao. 2013. “Poverty Impedes Cognitive Function.” *Science* 341 (6149): 976–80. https://doi.org/10.1126/science.1238041.

Mazhari, Tuba, and Graeme Atherton. 2021. “Students’ Financial Concerns in Higher Education.” *Higher Education Quarterly* 75 (1): 6–21. https://doi.org/10.1111/hequ.12267.

Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. 2014. “Sentiment Analysis Algorithms and Applications: A Survey.” *Ain Shams Engineering Journal* 5 (4): 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011.

Medvedev, Alexey N., Renaud Lambiotte, and Jean-Charles Delvenne. 2019. “The Anatomy of Reddit: An Overview of Academic Research.” In *Dynamics On and Of Complex Networks III*, edited by Fakhteh Ghanbarnejad, Rishiraj Saha Roy, Fariba Karimi, Jean-Charles Delvenne, and Bivas Mitra, 183–204. Springer Proceedings in Complexity. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-14683-2_9.

Pennebaker, James, Martha Francis, and Roger Booth. 1999. “Linguistic Inquiry and Word Count (LIWC),” January.

Pennebaker, James W. 1993. “Putting Stress into Words: Health, Linguistic, and Therapeutic Implications.” *Behaviour Research and Therapy* 31 (6): 539–48. https://doi.org/10.1016/0005-7967(93)90105-4.

Pennebaker, James W., Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. “The Development and Psychometric Properties of LIWC2015,” September. https://repositories.lib.utexas.edu/handle/2152/31333.

Pennebaker, James W, Cindy K Chung, Molly Ireland, Amy Gonzales, and Roger J Booth. 2007. “The Development and Psychometric Properties of LIWC2007,” 22.

Rajaraman, Anand, and Jeffrey David Ullman, eds. 2011. “Data Mining.” In *Mining of Massive Datasets*, 1–17. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139058452.002.

“Reddit.” 2022. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Reddit&oldid=1085020598.

Reddit. (2022). *Personal Finance Wiki* https://www.reddit.com/r/personalfinance/wiki/index

Schor, J.B. 1998. *The Overspent American*. Basic Books. https://research.tilburguniversity.edu/en/publications/15737094-9899-4273-9c79-c1f55dbea17c.

Schor, Juliet. 2008. *The Overworked American: The Unexpected Decline Of Leisure*. Basic Books.

Shen, Judy Hanwen, and Frank Rudzicz. 2017. “Detecting Anxiety through Reddit.” In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, 58–65. Vancouver, BC: Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3107.

Sievert, Carson, and Kenneth Shirley. 2014. “LDAvis: A Method for Visualizing and Interpreting

Topics." In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70. Baltimore, Maryland, USA: Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3110.

Stone, Philip J., Dexter C. Dunphy, and Marshall S. Smith. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. The General Inquirer: A Computer Approach to Content Analysis. Oxford, England: M.I.T. Press.

Taft, Marzieh, Zare Hosein, and Seyyed Mehrizi. 2013. "The Relation between Financial Literacy, Financial Wellbeing and Financial Concerns." *International Journal of Business and Management* 8 (May). https://doi.org/10.5539/ijbm.v8n11p63.

Tausczik, Yla R., and James W. Pennebaker. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29 (1): 24–54. https://doi.org/10.1177/0261927X09351676.

U.S. Department of Labor, Bureau of Labor Statistics, Consumer Expenditure Survey, Interview Survey, 2007