Title
Understanding Unconscious Bias by Large-scale Data Analysis

Author
Tang, Shiliang

Publication Date
2019

Peer reviewed|Thesis/dissertation

University of California
Santa Barbara

Understanding Unconscious Bias by Large-scale Data Analysis

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Computer Science

by

Shiliang Tang

Committee in charge:

Heather Zheng, Co-Chairperson
Ben Y. Zhao, Co-Chairperson
William Wang

September 2019

The Dissertation of

Shiliang Tang is approved:

William Wang

Ben Y. Zhao, Co-Chairperson

Heather Zheng, Co-Chairperson

May 2019

Understanding Unconscious Bias by Large-scale Data Analysis

To my family, friends, and loved ones

Acknowledgements

I would like to thank my advisors Professor Ben Y. Zhao and Professor Haitao Zheng for their tireless guidance throughout my PhD program. The were always super helpful on all aspects, both in research and in life. They always encourage me to be thoughtful, motivated, productive, and enthusiastic. I would also like to thank my committee Professor William Wang for his valuable feedback on my project and my PhD thesis.

I would also like to thank my collaborators for their hard work. They have always been helpful on making my research moving towards a optimistic direction. First, to the members of SAND Lab: Xinyi Zhang, Jenna Cryan, Xiaohan Zhao, Qingyun Liu, Gang Wang, Sixiong Shan, Divya Sambasivan, Yun Zhao, Megan McQueen. Second, to Scott Count and Apruv Jain at Microsoft Research, Ziming Wu and Professor Xiaojuan Ma in Hongkong University of Science and Technology, Professor Miriam Metzger from Communication Department at UC Santa Barbara.

I am also very thankful to the members in SAND Lab for discussing research ideas and making my life enjoyable. Apart from names I mentioned before, I would also like to mention Yanzi Zhu, Zhijing Li, Kevin Y. Yao, Zhujun Xiao, Huiying Li, Yuxin Chen, Emily Wilson, Zengbin Zhang, Yibo Zhu, Max Liu and Olivia Sturman.

Finally, I would like to express my spacial thanks my family, my friends, and my loved ones. They always show care and support unconditionally, which greatly encourage me to overcome any barriers throughout my PhD live.

Curriculum Vitæ
Shiliang Tang

Education

2014-2019	Ph.D. in Computer Science, University of California, Santa Barbara.
2010-2014	Bachelor of Science in Electronic Engineering, Tsinghua University, China.

Publications

Towards Understanding the Adoption and Social Experience of Digital Wallet Systems. **Shiliang Tang**, Ziming Wu, Xinyi Zhang, Gang Wang, Xiaojuan Ma, Haitao Zheng and Ben Y. Zhao. In *Proceedings of HICSS, 2019*

Penny Auctions are Predictable: Predicting and profiling user behavior on DealDash. Xinyi Zhang, Shawn Shan, **Shiliang Tang**, Haitao Zheng and Ben Y. Zhao. In *Proceedings of ACM Conference on Hypertext and Social Media (HT), 2018*

Gender Bias in the Job Market: a Longitudinal Analysis. **Shiliang Tang**, Xinyi Zhang, Jenna Cryan, Miriam Metzger, Haitao Zheng and Ben Y. Zhao. In *Proceedings of CSCW, 2018*

Echo Chambers in Investment Discussion Boards. **Shiliang Tang**, Qingyun Liu, Megan McQueen, Scott Counts, Apurv Jain, Haitao Zheng, and Ben Y. Zhao. In *Proceedings of ICWSM, 2017*

Cold Hard E-Cash: Friends and Vendors in the Venmo Digital Payments System. Xinyi Zhang, **Shiliang Tang**, Yun Zhao, Gang Wang, Haitao Zheng, and Ben Y. Zhao. In *Proceedings of ICWSM, 2017*

Unsupervised Clickstream Clustering For User Behavior Analysis. Gang Wang, Xinyi Zhang, **Shiliang Tang**, Haitao Zheng, Ben Y. Zhao. In *Proceedings of CHI, 2016*

Network Growth and Link Prediction Through an Empirical Lens. Qingyun Liu, **Shiliang Tang**, Xinyi Zhang, Xiaohan Zhao, Haitao Zheng, Ben Y. Zhao. In *Proceedings of IMC, 2016*

Abstract

Understanding Unconscious Bias by Large-scale Data Analysis

by

Shiliang Tang

Biases refer to disproportionate weight in favor of or against one thing, person, or group compared with another. Bias study has long been an important topic in psychology, sociology and behavioral economics. Over the years, people have observed large varieties of biases, proposed explanations to how they come, and identified how they impact our society.

A large amount of biases happens in an unconscious manner. They are built-in shortcuts in our brain that processes information automatically. Because of the unconsciousness nature of biases, studying biases usually comes from carefully designed controlled experiments to discover or explain unconscious biases. Over the past century, these experiments have identified over 200 different kinds of human biases and developed rich to explain the underlying mechanisms of biases.

However, because experiments happen in isolated environments with limited number of participants, they do not reflect these human biases in the wild. This task is hard in the years when the biases were first discovered when large-scale user behavior data are not largely available. With more and more user activities move online, gathering user behavior data are becoming more easier. These data provide us valuable opportunities for examining how human biases affect our society in the wild, and how biases affect our society when a large group of biased people freely interact with each other.

In this dissertation, we use empirical approaches to understand and measure human bias using large- scale datasets. We do not aim at identifying new types of biases or

measuring biases of individuals, but we focus on observing *aggregated* outcome of a group of biased people interacting with each other. This dissertation contains 4 studies strongly related to this topic.

The first two studies measure irrational behavior in the financial domain. We start with examining the quality of information and communication in online investment discussion boards. We show that positivity bias and skewed risk/reward assessments, exacerbated by the insular nature of the community and its social structure, contribute to underperforming investment advice and unnecessary trading. Discussion post sentiment has a negligible correlation with future stock market returns, but does have a positive correlation with trading volumes and volatility. Our trading simulations show that across different timeframes, this misinformation leads 50-70% of users to underperform the market average. We then examine the social structure in communities, and show that the majority of market sentiment is produced by a small number of community leaders, and that many members actively resist negative sentiment, thus minimizing viewpoint diversity.

Then we study the phenomenon of herd behavior of individual stocks. We hypothesize that in less mature markets, investors rely on common external inputs (*e.g.* technical analysis, or the generation of buy/sell "signals" using popular algorithms or software), resulting in herd behavior that moves the price of individual stocks. Our survey finds that Chinese investors in rely on technical analysis for investment decisions, compared to a minority of US investors. Next, using US markets as a baseline, we analyze two decades of historical price data on US and Chinese markets, and find significant support for the hypothesis that over-reliance on technical analysis has led to a "self-fulfilling prophecy" effect that makes prices of Chinese stocks much more predictable. Our trading simulation shows that by identifying and exploiting herd behavior, trading strategies based solely on technical analysis can dramatically outperform markets in China, while

severely underperforming in US markets.

The last two studies examine gender bias in different aspects. We start with study the effects of potentially gender-biased terminology in job listings, and their impact on job applicants, using a large historical corpus of 17 million listings on LinkedIn spanning 10 years. We develop algorithms to detect and quantify gender bias, validate them using external tools, and use them to quantify job listing bias over time. We then perform a user survey over two user populations to validate our findings and to quantify the end-to-end impact of such bias on applicant decisions. We show gender-bias has decreased significantly over the last 10 years. However, we find that impact of gender bias in listings is dwarfed by our respondents' inherent bias towards specific job types.

Following this study, we seek to systematically examine the problem of detecting gender stereotypes in natural language. We develop a gender stereotype lexicon that reflects the concept of gender stereotypes in the modern society, and compare the performance of traditional lexicon approach to an end-to-end deep learning approach on a large human labeled text corpus collected by us. We show that end-to-end approach significantly outperforms the lexicon approach, suggesting that in the future the widely used lexicon approach will be replaced.

In summary, we develop tools that are able to measure the human biases in large-scale, and perform large-scale measurements of aggregated behavior of human biases. We hope our work can shed light on a deeper understanding of human biases in the wild.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

People are not always objective. Over the years, researchers have discovered a large variety of human biases that shows what people understand or believe do not reflect what is actually happening in the real world. These biases can affect our behavior in different life aspects, including evaluating people, searching online, purchasing goods, selecting a place to live and choosing a career path.

However, it is well accepted by neural scientist that the majority of cognitive processing occurs in an unconscious way, because human brains can only actively process a very limited amount of information [1]. This also applies to human biases: often we do not control, or even are not aware of the existence of these biases. Sometimes, people's behavior can contradict their intention.

Because of the unconsciousness nature of biases, studying biases usually rely on carefully designed controlled experiments that observes people's behavior under different conditions. For example, psychologist Peter Wason invented the term "confirmation bias", by deigning an experiment that shows people only test their hypothesis in a one-sided way [2]. In the experiment, participants were told that (2,4,6) fits the rule that apply to a triple of numbers, and they could generate their own triples to find what is the

hidden rule. The result comes that the participants are extremely hard to find that the rule is "any ascending sequence", because they often found a more specific rule and only test positive examples. In another experiment that seeks to demonstrate unconscious gender bias, participants are asked to role-play as hiring manager to go over resumes of applicants. It has been find that participants displayed a strong tendency to choose male candidates, even if a female candidate appears as a slightly better performer [3].

Using experiments, previous studies have identified over 200 different kinds of human biases. Rich theories have been developed to explain what is the underlying mechanisms of biases. For example, one of the most popular explanation around human biases states that biases come from "mental shortcuts" in our brain that rely on a limited number of heuristic principles to reduce the complexity of tasks [4]. In a different explanation, bias is a evolved function for avoiding costly false negatives (failing to take an action that would have been better to take) [5]. Although there is no consensus on which explanation is "correct", they all successfully predict a certain amount of biased behavior.

Although these experiments are powerful in identifying or explaining human biases, they can not solve the problem on how does these humane biases perform in the wild. The experiments usually happens in a isolated environment with limited number of participants, so that there is lack of observation of how biases affect our society when a large group of people freely interact with each other. This task is hard in the years when the biases were first discovered, at which time large-scale user behavior data are not largely available. For example, confirmation bias was identified in 1960s, but it is until 2000s that research identified echo-chamber effect, a direct result of confirmation bias when it is applied to a group of people in a community. With more and more user activities move from offline to online, gathering user behavior data are becoming easier. These data provide us valuable opportunities for examining how human biases affect our society in the wild.

In this dissertation, we use empirical approaches to detect and understand human bias using large-scale datasets. We do not aim at identifying new types of biases or measuring biases of individuals, but we focus on observing *aggregated* outcome of a group of biased people interacting with each other.

My dissertation contains 4 studies strongly related to large-scale measurement of 3 types of unconscious bias: confirmation bias, herd behavior and gender bias, as introduced in Chapter 2. The first 2 studies demonstrate aggregated outcome of a group of biased people in financial domain. We start with examining echo chamber effect in the investment discussion boards, as echo chamber effect is a unconscious exercise of confirmation bias. We take user discussion threads over large online forums, to demonstrate how echo chamber effect affects quality of user generated content (Chapter 3). Then we look at how herding behavior of individual investors drive market change, as herding behavior has repeatedly found in irrational individual stock investors (Chapter 4). The last 2 studies examine gender bias in different aspects. We look at how gender bias in the job advertisements affect gender inequality of the job market. We download millions of job advertisements over 10 years of history, and demonstrate the changing trend of gender biased language and their effect on potential job applicants (Chapter 5). Following this study, we develop and evaluate tools for detecting languages that reflect gender stereotypes. We gather first-of-a-kind data about gender stereotype language, and systematically compare pros and cons of two different approaches for this task (Chapter 6).

In the following, we briefly introduce the work included in this dissertation.

1.1 Echo Chamber Effect in Investment Discussion Boards

Today, online communities are getting more and more popular. People search, consume, and share large amount of information through discussions in online communities. Being a place where people can freely exchange information and opinions, online community is vulnerable to echo chamber effect: like-minded people voluntarily "herd" together, reinforcing existing biases of these people. It has been found that echo chamber is affecting users in traditional online social network, like Facebook or Twitter [6, 7]. In online communities with echo chamber, discussions are likely to be isolated from the real world, and opinions are usually polarized and monotone [8].

In this work, we examine whether echo chamber also presents in investment discussion boards. Investment discussion boards are popular online forums where investors share and discuss their opinions on investments of mutual interests. These forums support millions of investors, and are often one of the top results from search engines in response to queries for stock symbols. They are very influential for investors looking for financial advice. The consequence of an echo chamber on investment discussion boards would be that investors irrationally follow the investment advice which leads to direct financial loss.

We try to determine the presence of echo chambers in investment discussion boards using an empirical approach driven by data from two of the largest stock discussion boards, Yahoo Stock Message Boards and InvestorHub (iHub). We download message logs for discussion boards on both sites covering publicly listed stocks of US-based companies, totaling 34 million message posts across nearly 13,000 stocks. We apply a customized sentiment analysis tool on messages in the investment discussion boards, and put users into behavior clusters based on their patterns of interaction.

Our analysis demonstrates that there is near zero correlation between discussion sentiment and short-term price movements of a stock. Trading strategies based on sentiment significant underperform market baseline. Furthermore, discussion threads on most boards might be controlled by a very small group of users. Negative sentiment drops significantly as threads are taken over by large groups of "reacting users and followers" whose posts reinforce each other and the overall sense of positivity in these boards.

1.2 Herding Behavior in Stock Market

Stock market is a social computational tool, such that the market price is driven by a complex interaction of investor buy and sell decisions. Individual investors in the stock market are not always rational, as it has been repeatedly found that people tend to adopt the trading strategy of the majority [9]. As a result, the strategy adopted by the majority would in return drive the up-and-downs in the market. This is especially true if the market is immature: when high quality fundamental financial data is not publicly available.

In this study, we study the phenomenon of herd behavior on each individual stocks. We hypothesize that in less mature stock market in China, individual investors need to make trading decisions based on patterns in price change, which is usually termed as *technical analysis*. In a market where a large amount of people is using technical analysis, the market is likely to shift in a pattern that are consistent with the patterns used in the technical analysis. Such shift may yield positive return for these traders in short term, but it can cause market bubbles and increase the risk of market crash in a longer term.

To study if individual investors in immature stock market use technical analysis in a large scale, we conducted a comparative study that examine individual behavior and market outcomes in Chinese and US stock market. We first deploy a user survey, and

find that while few US-based investors rely on technical analysis for investment decisions, all China-based investors in our study used technical analysis on a regular basis. Chinese investors also rely on media sources for making trading decisions. Then, we analyze over two decades of stock data in US and Chinese markets, and find that technical analysis tools are significantly more accurate when applied to stocks in the Chinese market than those in the US market. The most popular tools reported in our survey also yield the highest accuracy in predicting market change. A trading strategy solely based on technical analysis can significantly outperform market baseline. In total, our study provide strong evidences that investors in China collectively apply technical analysis which drives significant market change.

1.3 Longitudinal Measurement of Gender Bias in the Job Market

Gender inequality persists in workplace today [10, 11]. Women still earn much less than men [12], and there are significantly fewer women in male dominated positions, across both industry sectors (*e.g.,* technology [13]), and job types (*e.g.,* corporate executives [14]).

There are two well-documented reasons for gender disparity in the job market. First, people have stereotypes of genders and occupations. People are most likely to seek out occupations that are compatible with one's sense of self [15], and people assume stereotypes associated with the gender of a worker must correlate with the requirements of their occupation [16]. Second, institutional gender discrimination exists in the job market. For example, in application for jobs in restaurants, female applicants were significantly less likely to get an offer from high-end restaurants compare to comparably

matched men [17].

More recently, it has been found that words in job advertisements also have some impact on gender inequality in the job market. In a lab experiment evaluating language in job advertisements, using masculine words (*e.g.,* "strong", "competitive") may cause female applicants unwilling to apply for the job [18]. Base on these results, online services like Unitive and Textio developed algorithms to detect such language that may cause gender bias.

The goal of this work is to understand the role that job postings play in introducing or exacerbating gender imbalance in the workplace. We perform a study with two components. First, we analyze 17 million job posts collected over the last 10 years. We obtain this dataset of job posts from LinkedIn, the largest professional networking site online with 500 million users. We develop two scalable algorithms that match the same metrics as online services that evaluate job postings for gender bias, and apply our algorithm to the LinkedIn dataset to quantify gender bias in the whole market, specific sectors, and its changes over time. Second, we ran a controlled experiment to quantify the effect of the language of the job postings. We create two versions of a small sample of job postings. In one version all the masculine and feminine words are replaced by gender neutral words without significantly changing the meaning of the job postings. We show these job postings to two user populations and asked detailed questions about their perception about the job postings.

Our empirical data analysis demonstrates that the job market is dominated by masculine language over the past decade, and the bias is dropping. However, bias level vary a lot across different job sectors, employment types and seniority levels. Significant bias still remains. Our controlled experiment shows that replacement of masculine and feminine words only have low levels of impact on user decisions to apply. User stereotypes play a more significant role than the gender biased language we observed.

1.4 Detect Gender Stereotype in Natural Language

Not only job advertisements, gender stereotypes are persistent in other types of articles, such as notable people biographical page [19], recommendation letters [20], fictions [21] and movie dialogs [22]. The prevalence of gender stereotype in languages may reinforce the concept of gender stereotypes like aggressive men and domestic women.

Although gender stereotype are frequently observed, the detection mechanisms are not fully developed. Previous works approaching this problem using a lexicon based approach. Specifically, a list of words that reflect masculinity or femininity are compiled, and the frequency of these words occur in an article indicate the strength of gender stereotype. However, it was found that items captured by existing lexicon are less endorsed by women in recent years [23, 24] Although this approach is applied as an industry state-of-the-art for detection gender biased language in job advertisements [18, 25], its performance remains unclear.

With the wide availability of large-scale text data and development of natural language processing techniques, an end-to-end approach starts to outperform word lexicon based approach. Instead of pre-compile a word lexicon, the end-to-end approach trains a neural network model that produce the desired output using raw text as input. The end-to-end approach has shown great success in tasks such as sentiment analysis **??** or hate speech detection [26], which were traditionally tackled by word lexicon approach [27, 28].

In this work, we seek to develop a gender stereotype lexicon that reflect gender stereotypes of the modern society, and understand if the end-to-end approach can outperform traditional lexicon approach. For word lexicon approach, we compile a gender stereotype lexicon that contains over 10 thousand verbs and adjectives. Each word has a score that indicates whether the word is masculine or feminine. We then collect a large number of human labeled articles that are consistent with or contradict gender stereotypes. We

apply the lexicon approach and the end-to-end approach on our dataset. Our results indicate that the end-to-end approach significantly outperforms lexicon approach, even with limited amount of training data. The end-to-end approach is better at inferring semantics of sentences, as it performs well in cases when the lexicon approach fails to understand subtle language phenomenon (*e.g.,* negation).

Chapter 2

Background

2.1 A Introduction to Cognitive Bias

Studying human biases has a long history dating back to 1950s. Biases refer to disproportionate weight in favor of or against one thing, person, or group compared with another. Bias study has long been an important topic in psychology, sociology and behavioral economics. Over the years, people have observed a large varieties of biases, proposed explanations to how they come, and identified how they impact our society.

Cognitive bias is one of the most studied classes of bias. It refers to cases in which human cognition produces representations that are systematically distorted compare to objective reality [5]. For example, in judging the sequence of coin flips, people assessed the sequence HTTHTH to be more likely than the sequence HHHTTT, which violates simple probability theory [29].

Cognitive bias arises from various processes. The most commonly invoked process is *heuristics*. Because human brain needs to process massive information from the outside world, the brain develops a number of "mental shortcuts" that rely on a limited number of heuristic principles to reduce the complexity of tasks [4]. The heuristics are appealing in

that the cost to use a sophisticated strategy usually do not compensate for the increased accuracy [30]. Other than heuristics, previous studies also proposed that cognitive bias is a evolved function for avoiding costly false negatives (failing to take an action that would have been better to take) [5].

Given that there are hundreds of different kinds of cognitive bias, we are not listing or introducing all of them. In the rest of this chapter, we briefly introduce a few variations of cognitive biases that are directly related to the studies in this dissertation.

2.2 Confirmation Bias and Echo-chamber Effect

Confirmation bias is a type of cognitive bias. It refers to the tendency for seeking or interpreting of evidence in ways that confirm existing beliefs, expectations, or a hypothesis in hand [31]. Confirmation bias affect how people search, interpret and memorize information. For instance, experiments have found repeatedly that people test hypothesis in a one-sided way, by only searching for evidence consistent with their current hypothesis [31]. People with opposite view points feel their opinions to be more convincing after reading the same material, because they only focused on evidences supporting their viewpoint and disregarded anything contrary [32]. Moreover, people selectively remember facts that are consistent with their prior expectations [33].

Echo-chamber effect is a unconscious exercise of confirmation bias. It usually happens on open platform when groups of people can freely access, search and exchange information, like online forums or social network. Echo chamber effect describes a social discourse environment in which ideas tend to be mutually reinforced by a like-minded group of people [8], and have been found to evolve organically in a variety of contexts. For instance, in a study of blogs, agreement in comments outweighed disagreement by a ratio of three to one [34]. Analyses of controversial debates on Twitter have shown

that while discussions with different-minded individuals can strengthen out-group affiliations, discussion among like- minded individuals only strengthens group identity [7]. Further, people are more actively engaged in their online communities if they encounter less disagreement, especially in talking about political topics [35].

Confirmation bias and echo-chamber effect can lead to negative consequences. In political sphere, group of people becomes insulated from diverse perspectives, because the massive number of online information sources allows people to gravitate to politically opinion-confirming Internet news [36]. In finance, confirmation bias can lead investors irrationally follow the advice and behavior of others while ignoring signals of losing money [37]. Even in scientific research, researchers tend to evaluate results qualitatively higher when they conformed to their own prior expectations. Biased researchers give little serious thought to revising nonconforming results [38].

2.3 Herd Behavior

People have tendency to follow others, which is usually referred to as the "bandwagon effect". The bandwagon effect states that the rate of uptake of beliefs, ideas and trends increases the more that they have already been adopted by others [39, 40]. For example, in a lab experiment, people appreciate news stories more if they are told that the stories are also selected by "other users", compare to the same stories but labeled as selected by new editors or computers [41]. The bandwagon effect is also observed in political elections, where voters favor a party that is doing well in the polls [42]. The bandwagon effect is part of the large classes of cognitive biases that influence judgments and decisions of people.

Bandwagon effect causes *herd behavior*, which refers to how individuals in a group can behave collectively without centralized direction. Herd behavior is repeatedly observed

in making investment decisions. In financial markets, herd behavior is often cited as irrational behavior that spreads through entire markets during periods of market extremes, either during "bubbles" created by greed, or market crashes driven by collective fear [9]. A well-cited user survey found herding behavior spread through word-of-mouth between institutional investors [43]. Herd behavior is also used by marketing. For example, star ratings and sales volume are shown to be great incentives that cause purchase actions in large numbers of people [44].

2.4 Stereotype and Implicit Bias

A *stereotype* is an over-generalized belief about a particular category of people. For example, according to gender stereotype, women should display communal traits (*e.g.,* , nice, caring, warm) and men should display agentic traits (*e.g.,* , assertive, competent, effective) [45, 46]. Similarly, in racial stereotypes, black people are athletic, musical and poor, while white people are intelligent, ambitious and greedy [47].

Stereotypes come from a mixture of cognitive bias and social factors. For cognitive bias, it has been found that people have the tendency to ascribe a person's behavior to disposition or personality, and underestimate the role of external factors [48]. Also, people are likely to overestimate the frequency of co-occurred low frequency events, *e.g.,* black people and crime [49]. For social factors, social groups are motivated to behave in certain ways, and the stereotypes reflect those behaviors. For example, gender stereotypes originate in the gender-typical social roles and sexual division of labor. Women are more likely than men to be homemakers, and less likely to be employed in the paid workforce. As people observe sex differences in distribution of occupational roles, people align gender with corresponding occupational characteristics [45].

Adoption of stereotypes are usually unconscious. Termed as *implicit bias* or *implicit*

stereotype, it is the stereotype that results from subtle cognitive process that often operate at a level below conscious awareness [50]. Implicit stereotype stems from part of people's mental system. According to a well accepted framework of cognitive functioning [51], human mental system contains two parts: part 1 handles cognition that occurs outside of conscious awareness which operates automatically and fast, part 2 is conscious processing that requires deliberate concentration. It is will accepted that conscious information processing capacity is limited, and the majority of cognitive processing occurs in unconscious part [1].

Implicit stereotype is developed over time with accumulation of personal experience. For example, children in fourth to fifth grade greatly demonstrate similar implicit stereotype of their parents' [52]. In another experiment, negative emotions like anger or disgust triggered by other social groups may activate implicit bias towards that social group [53]. Even for people who do not endorse stereotypes, understanding of the stereotypes can foster implicit stereotype [54].

As implicit stereotypes are unconscious, observation and measurement of implicit stereotypes are usually through *priming*: response to a target (*e.g.,* bread) is faster when it is preceded by a related prime (*e.g.,* butter) compare to an unrelated prime (*e.g.,* dog). In a lab experiment about gender stereotypes, seeing words related to stereotypes (*e.g.,* nurse, doctor) followed by target pronouns (*e.g.,* she, he) can trigger faster response on classifying the pronouns into gender [55]. Based on this phenomena, Implicit Association Test [56] is developed as a standard test of implicit bias. Implicit Association Test captures the difference in response times when participants are asked to pair two concepts that they find similar, in contrast to two concepts they find different. For example, participants pair flowers to "pleasant" faster than pair them to "unpleasant". The test can be applied to detection of implicit bias in a large variety of topics, including gender, race, career, science and disability.

Implicit stereotype can harm effectiveness of professional judgments. In one experiment, police office shoot armed black target faster than armed white target [57]. Similarly, likelihood of being sentenced to death is influenced by how much the black defendant's appearance is perceived as stereotypically black [58]. In a lab experiment that simulated a hiring decision process, male participants displayed a strong tendency to choose male candidates, even if a female candidate appears as a slightly better performer [3]. In another field study, comparably matched men and women are sent to apply for jobs in restaurants, and the study found female applicants were significantly less likely to get an offer from high-end restaurants [17].

Chapter 3

Echo Chambers in Investment Discussion Boards

3.1 Introduction

Investment discussion boards are popular online forums where communities of individual investors share and discuss their opinions on investments (commonly stocks) of mutual interest. These forums support millions of investors, and are often amongst the first results returned by search engines in response to queries for stock symbols. As such, they can be very influential for investors looking for financial advice.

But are these forums providing a useful service to their millions of users? What is the quality of discourse, and how effective are they as mechanisms of disseminating useful financial information? As platforms for social discourse where participation is self-selected, there is a distinct possibility of naturally forming *echo chambers* that can dramatically affect the quality of information exchanged on these forums. Echo chambers are discussion environments where ideas tend to be mutually reinforced by like-minded people, often contrary to external inputs [8], and have been found to evolve organically

in a variety of contexts, including political forums, blogs, and social networks such as Twitter [34, 7, 6].

While there could be rational reasons to "herd" when the community knows something the individual does not, our results indicate this is not the case. Instead, by reinforcing existing biases and insulating users from objective facts, echo chambers are likely to induce poor investment decisions with potentially severe consequences. In the context of finance theory, echo chambers could play a role in producing *noise traders*, investors who make decisions driven by biases and emotions that push prices of investments away from their true value. In aggregate, noise traders can negatively impact markets by increasing price volatility and contributing to price bubbles [59, 60].

In this paper, we try to determine the presence of echo chambers in investment discussion boards using an empirical approach driven by data from two of the largest stock discussion boards, Yahoo Stock Message Boards and InvestorHub (iHub), that together claim hundreds of millions of registered users and tens of millions of visitors per month. We download message logs for discussion boards on both sites covering publicly listed stocks of US-based companies, totaling 34 million message posts across nearly 13,000 boards. Our work focuses on three questions:

- Does sentiment in stock discussion boards reflect an "echo chamber" effect, where sentiment decouples from external information, *e.g.* stock price fluctuations? To what extent, if any, are these effects correlated with level of user attention, participation, and bias in the forum?
- What is the internal structure of communication inside these boards, and what roles do individuals play in the depth and length of specific discussions? Is there any relationship between communication across specific groups and insulation of sentiment against external factors?

- What empirical impact do discussions on investment boards have on trading performance? Can investors rely on discussion sentiment to outperform stock market indices? Does investment board activity correlate with other properties of stock trading such as volume and volatility?

We summarize our contributions below.

First, we prune our data to identify message boards that focus on publicly traded companies listed on US exchanges. We analyze gross message volume for both Yahoo and iHub, and show that the highest levels of user activity often reside on small market cap stocks with higher risk and volatility.

Second, using customized sentiment analysis tools, we extract aggregate sentiment out of daily discussions from each of the message boards in our analysis. We demonstrate that across different platforms, parameters, and metrics, there is near zero correlation between discussion sentiment and short-term price movements of a stock. Regardless of how a stock performed, sentiment in discussion boards is almost uniformly positive and upbeat.

To confirm the lack of value in board sentiment, we run detailed trading simulations to quantify the predictive quality of discussion board sentiment. We find that trading based on discussion board sentiment significantly underperforms market indices. While we cannot know to what extent discussion boards influence active (or passive) users, the general lack of diversity of opinions hurts the quality of the advice overall.

Third, we show that while board sentiment does not predict future stock performance, board activity as measured by the number of posts, does correlate with future stock trading volume and volatility. Importantly this correlation was strongest for stocks with the most attention by board members. Together with the lack of effect for board sentiment, this suggests that activity in the discussion boards signals misguided market activity and potentially is a source of market noise. These results are broadly consistent with financial

economics literature with studies done on smaller samples (see [61] for a summary).

Finally, to better understand *why* these boards are poor sources of information, we explore their echo chamber-like qualities and show they are insular and actively reject disliked viewpoints. We use unsupervised learning techniques to identify clusters of users with similar communication patterns. We find that discussion threads on most boards might be controlled by a very small group of users, and negative sentiment drops significantly as threads are taken over by large groups of "reacting users and followers" whose posts reinforce each other and the overall sense of positivity in these boards. Using a small dictionary of keywords expressing personal animosity, we find that posts with negative stock sentiment tend to evoke hostile responses from the community, perhaps explaining why users who post the most negative content tend to be outside of the dominant user groups.

3.2 Noise Traders and Echo Chambers

We begin with an introduction to the phenomenon of noise trading in the context of finance theory, and discuss the possible role that echo chambers play in financial markets.

Noise Traders. In classical finance theory, rational investors should only trade in ways consistent with the objective probabilities of the state of the economy. Increasingly, the financial literature recognizes that the average investor is not the perfectly rational investor of theory [60]. In this paper we investigate how the social mechanism of message boards may introduce noise and hence contribute to the lack of useful economic information (trading signals) and potentially result in irrational or noise trading.

We thus refer to the trading actions of individuals based on noise as noise trading, consistent with literature that defines noise traders as traders having imperfect or irrational expectations [62, 63, 64], some of which may arise from behavioral biases [65, 66]. The

role noise traders play is a source of debate, where some [62] argue that they provide beneficial liquidity, others argue that they can exert influence and drive the prices away from economic fundamentals and render the market inefficient [60]. One of the insights from the efficient markets hypothesis [67] is that new information should be unforecastable. Thus if message boards truly were places to collect valuable information, they should uncover opposing views with equal frequency. However, as we show, these boards appear to be influenced by social constructs that generate heavily biased information.

Messages Boards as Echo Chambers. Message boards and other online forums support discourse at scale, in both range of topics discussed and the number of people involved. Since participation is largely self-selected and often in the context of a social network, the emergence of echo chambers is a distinct possibility. Here we consider an echo chamber to be a social discourse environment in which ideas tend to be mutually reinforced by a like-minded group of people [8]. For instance, in a study of blogs, agreement in comments outweighed disagreement by a ratio of three to one [34]. Further, it has been shown that people are more actively engaged in their online communities if they encounter less disagreement [35]. More recently, Facebook has been examined for echo chamber effects, to mixed results. The Facebook newsfeed is slightly more aligned with an individual's own ideology, but a notable minority of friends and news feed content represents opposing views [6].

Given evidence of echo chambers online, what are the consequences? The primary concern is that a group of people becomes insulated from diverse perspectives. There is evidence from the political sphere, that online political forums can resemble echo chambers, in part because the massive number of online information sources allows people to gravitate to politically opinion-confirming Internet news [8, 36]. In addition to selective information exposure, there are network effects where online networks of those with more extreme views tend to be more homophilous [68].

Message Boards	# of Boards	# of Messages	# of Users	Data Since
iHub	6,839	5,329,547	56,421	03/25/2000
Yahoo	6,135	28,845,907	1,494,769	02/13/1996

Table 3.1: Basic statistics of crawled Yahoo and iHub boards.

Investments and Echo Chambers. Tying this together, in a financial investment setting, the consequences of an echo chamber would be that investors would irrationally follow the investment advice and behavior of others in financial discussion boards insulated from external market news. In other words, echo chambers in these online forums could give rise to the aforementioned noise traders. For individuals, this could lead to direct financial losses through poor investments. For the market as a whole, this can lead to inefficiencies if enough people make poor investment decisions.

3.3 Data Collection and Preliminary Analysis

Investors Hub (iHub) and Yahoo Message Boards are the most popular online financial forums for individual investors. These financial forums are built as platforms for individual investors to converse and share insights related to stock picks and trading strategies. Content is usually organized as individual message boards, each associated with a specific stock or fund.

iHub. iHub has been online since early 2000, and claims to have 562M registered users, and 122M messages on 25,000 boards[1]. Each iHub board is moderated for content, and registered users can start new message threads, reply to messages, and *follow* users to receive automatic updates.

In March 2016, we crawled all posts under iHub boards associated with stocks listed in US markets. Our dataset has a total of 5,329,547 messages posted by 56,421 users across

[1] `http://investorshub.advfn.com/boards/about.aspx`.

Dataset	Exchange	# of Boards	# of Boards with Price
iHub	NASDAQ	2018	2014(99.8%)
	NYSE	1765	1757(99.5%)
	AMEX	214	213(99.5%)
	Other	2842	1471(51.8%)
Yahoo	NASDAQ	2608	2526(96.9%)
	NYSE	2404	2401(99.9%)
	AMEX	293	292(99.7%)

Table 3.2: Breakdown of both datasets in stock exchanges and stock price coverage. For Yahoo, we only consider boards with complete post history.

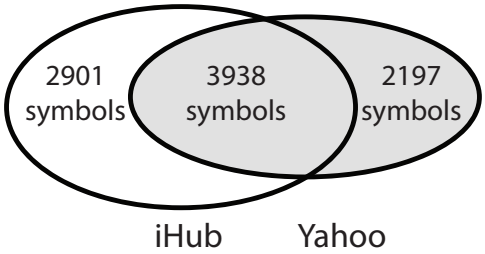
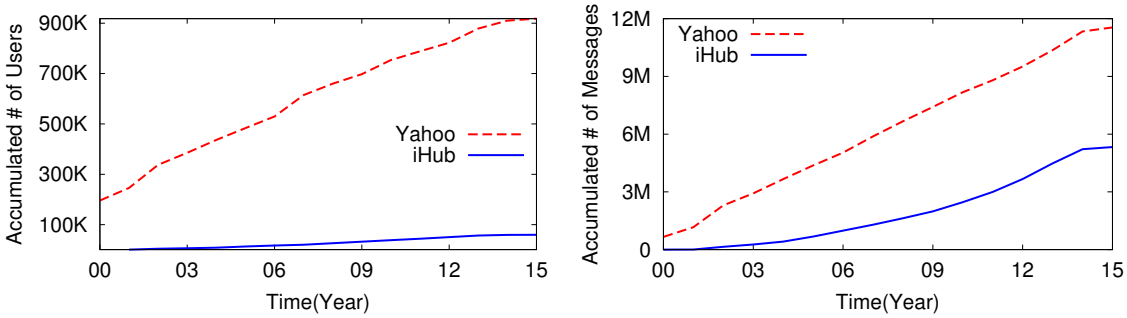


Figure 3.1: Stock Symbols extracted from iHub and Yahoo.

6,839 boards since March 25, 2000 (Table 4.1). Each message includes a message ID, user ID, timestamp, message text and an optional message ID of the message to which it is a reply. Each iHub board is associated with a ticker symbol, which is an abbreviation used to uniquely identify a particular stock. Among them, 60% are valid stock symbols listed in the top 3 US stock exchanges in the U.S.: New York Exchange (NYSE), NASDAQ, and American Stock Exchange (AMEX)[2]. The remaining 40% of symbols include names of non-listed companies, delisted stocks, stocks on international exchanges, small stocks sold on Over-The-Counter markets (OTC), stock funds and ETFs, stock options, or user-created names. Table 3.2 shows the number of boards associated with stocks in each major exchange.

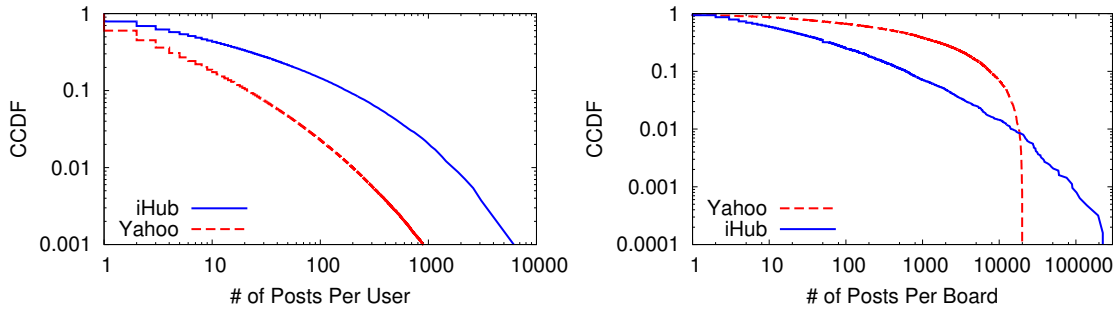
Yahoo Message Boards. Yahoo Message Boards are part of Yahoo! Finance, the most popular financial news website. Started in 1994, Yahoo! Finance claims 70M

[2]For stocks listed on multiple exchanges, we only count it once.



(a) User count (b) Post count

Figure 3.2: Total number of users and posts over time.



(a) Over Users (b) Over Boards

Figure 3.3: Distribution of posts over users and over message boards.

unique visitors each month. Yahoo Finance provides a webpage for each stock or fund, which links to its own message board. Message boards are created by default for each ticker symbol, even for non-tradable, non-stock entities like corporate debt. Users can post messages and replies on boards, and all messages in Yahoo! Message boards are public.

We used a list of publicly traded stock symbols to look up message boards for US-based stocks. We downloaded a total of 6,135 message boards, 28,845,907 messages by 1,494,769 users since Feb. 13, 1996 (Table 4.1). Of these boards, we retrieved all posts from 5,305 (86%) of them. The remaining boards had more content than what Yahoo Finance displays (a maximum of 1002 display pages). For consistency, we restrict analysis

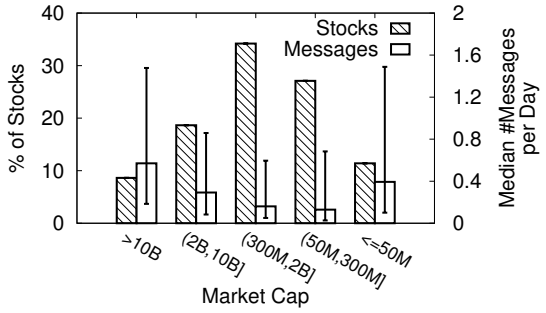


Figure 3.4: Distribution of Yahoo boards and average messages by market cap.

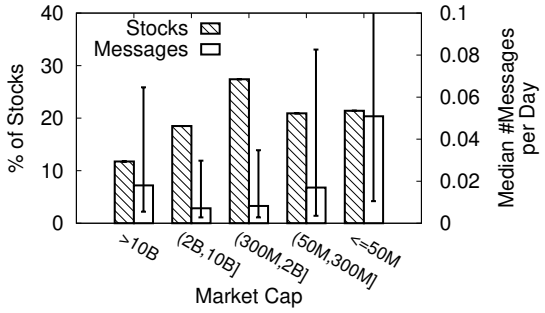


Figure 3.5: Distribution of iHub boards and average messages by market cap.

and simulation in the rest of this paper to the 5,305 fully crawled message boards. Board coverage from Yahoo boards is also listed in Table 3.2.

We compare symbols between Yahoo and iHub datasets in Figure 3.1. Most or all US-based stock symbols are found in both datasets. iHub-only symbols (2,901) are mostly invalid symbols, de-listed stocks, or small stocks traded by the OTC Market Group, while Yahoo-only symbols (2,197) are generally empty message boards associated with corporate debt or foreign stocks. Through manual inspection of these Yahoo-only boards, the large majority of posted messages seem to be spam or fake messages generated by automated scripts.

Finally, we used the Yahoo! Finance open API[3] to crawl historical stock prices for each stock symbol, including daily closing price and daily volume (# of shares traded). We then adjusted closing prices and volume to account for stock splits and dividends in order to finalize a set of complete daily price histories for all stocks.

3.3.1 Basic Analysis

Users, Posts and Distribution Across Boards. We show in Figure 3.2 the accumulated growth of users and posts over time for both datasets. While both are

[3] `http://finance.yahoo.com/`

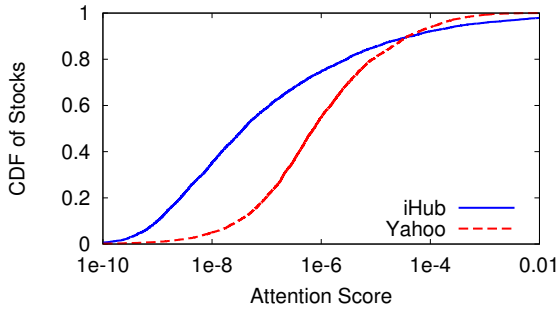


Figure 3.6: Distribution of stock attention scores for both datasets.

growing steadily in users and posts since 2000, we see that Yahoo boards have much faster growth in users. Yahoo's user count might benefits from the popularity of Yahoo Finance.

We plot the distribution of posts over users and boards in Figure 3.3. The distribution shown is using the Complementary Cumulative Distribution Function (CCDF). Both distributions are highly skewed. Roughly 80% of users on iHub and 95% of users on Yahoo post less than 50 posts, and a significant number of users do not post any messages. We see similar trends across message boards, where 80% of iHub boards and 40% of Yahoo boards have less than 200 posts over their lifetime. Relatively speaking, iHub's user population, while smaller, tends to generate more messages than Yahoo users. This is likely because discussion board is iHub's primary function, while message boards are only a small component of Yahoo Finance.

Distribution of Market Cap. *Market cap.*, short for market capitalization, is the total dollar market value of all of a company's outstanding shares. It effectively captures the size of a company, and is a critical part of any fundamental method for computing the valuation of a company and its shares. Market cap is also heavily associated with a stock's risk/reward trade-off. Generally speaking, large cap companies tend to be lower risk and give lower returns compared to companies with smaller market cap.

We use a variant of the categories defined in [69] and divide companies into 5 different categories using their market cap on April 5, 2016, as reported by Yahoo! Finance. Figure 3.4 uses a dual Y-axis graph to show the distribution of companies with Yahoo message boards sorted by market cap, and the median posts per day (along with 25th and 75th percentiles). Clearly, most of Yahoo boards belong to mid-cap sized companies, but boards with the heaviest post traffic are in the highly speculative, nano cap (≤50M) companies and the large cap companies. In contrast, as shown in Figure 3.5, iHub's message boards are more evenly distributed by company size, but iHub's heaviest traffic also comes from the small set of speculative, nano cap (≤$50M) companies.

User Attention. Finally, we propose a new metric called *attention score* to capture the level of user attention received by a stock. We define it as the number of total posts of a stock divided by its market cap. The goal is to normalize for the impact of proportionally larger number of investors in the user population who own a given large cap stocks, *e.g.* Apple. A higher score means that users are more active in talking about the stock. We plot the Cumulative Distribution Function (CDF) for message boards in both datasets in Figure 3.6, and note that while most boards in Yahoo have generally higher attention scores, the most active message boards at Yahoo are less active than those at iHub (Figure 3.3b). This is likely due to the extremely high activity on microcap stocks in iHub (market cap <$50M in Figure 3.5).

3.4 Sentiment Extraction

To measure the value of user posts on guiding investment strategies, we perform sentiment analysis on online discussion boards in the context of stock market, which explores interactions between media content and stock market activity. Previous work focused primarily on extracting content from news media [70], general social media [71, 72]

and finance-specific crowdsourced services [73].

Our first step is to develop reliable tools to interpret the sentiment of posts. We employ different approaches for the Yahoo and iHub datasets, due to differences in various properties of the two message boards. The Yahoo data has a subset of posts with sentiment naturally labeled by users themselves. We use these labeled posts as a training set, and apply a standard supervised learning approach, which is *classification based*. Unfortunately, iHub does not provide such labeling. Manually labeling a subset of posts that are large enough for classification could be both errant and time-consuming. To deal with this, we adopt a *keyword dictionary based* method. Our validation results show these tools achieve an accuracy of 72.1% for Yahoo and 81.0% for iHub, which are on par with or significantly better than popular sentiment analysis tools [74, 75].

3.4.1 Classification Approach (for Yahoo)

A classification approach trains a machine learning classifier on labeled posts, where labels are sentiment of posts, and features are extracted based on content of posts. The classifier is then applied to unlabeled posts to recover their sentiment. Roughly 5% of Yahoo messages already have sentiment labeled by users themselves, as one of the following: "Strong Buy", "Strong Sell", "Buy", "Sell" and "Hold". We regard both "Buy" and "Hold" as positive sentiment, while "Sell" as negative. Holding a stock is positive, because it means the owner is still expecting the stock price go up in the future. Out of these 5% of messages, we have 17.2% negative posts and 82.8% positive posts.

To generate features, we follow a supervised machine learning method from [73], where unigrams are regarded as features. Unigrams of a post are the numbers of occurrences of each unique word in the post. We also remove stop words, stock symbols and urls from messages, and exclude infrequent unigrams that occur less than 700 times[4]

[4]We tried thresholds from 100 to 2000 and found similar results.

over all messages. We apply Naive Bayes, Random Forest, Supported Vector Machine (SVM), Decision Trees and Logistic Regression from the standard machine learning package Scikit-learn [76].

We run 10-fold cross validation on the labeled data, and find that SVM performs best. When evaluating classification-based accuracy, we found the imbalance in training data, *i.e.*, fewer negative posts than positive posts, contributed to unacceptable classification errors. Specifically, precision for predicting positives and negatives is 86% vs. 56%, with an even bigger gap in recall at 97% vs. 21% respectively. we apply undersampling to build our training dataset, keeping all negative posts while varying the number of positive posts [77]. We vary the under-sampling ratio (number of positive posts : number of negative posts), from (1:1.0) to (1:4.8), and find 2.0 gives the best results, with an overall accuracy of 72.1%, 75% precision and 86% recall for positive predictions, 61% precision and 43% recall for negative predictions.

3.4.2 Keyword Based Approach (for iHub)

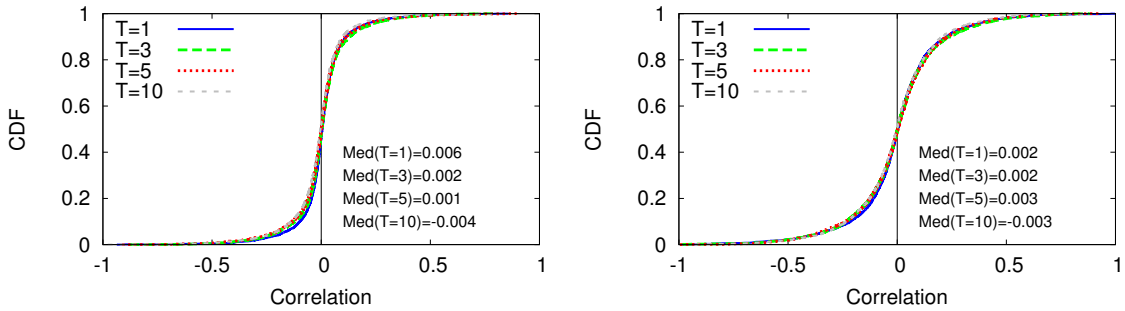
A keyword-based approach requires a dictionary of positive and negative keywords, with the resulting sentiment of a post extracted by counting the number of positive or negative keywords in the post.

The dictionary is generated by analyzing content of a small number of labeled posts. To generate such labeled posts for iHub, we randomly sampled 4,500 (about 0.1%) posts across all boards, and asked three native speakers in English to manually label them with one of the following: "Buy", "Sell" and "Hold". Note here we take a conservative approach in labeling, which means we only consider sentiment within the single post itself, without recovering its context in the whole message thread. This might introduces noise in sentiment extraction. Like Yahoo, we regard both "Buy" and "Hold" as positive

sentiment while "Sell" as negative. In total, we have 96% positives and 4% negatives for these 4500 posts. Next, we extract keywords from the newly labeled posts and build a sentiment dictionary to identify positive and negative keywords [73]. Specifically, we extract unigrams from labeled posts, separate unigrams of positive posts from those of negative posts, and rank them in descending order of frequency. Then for the top ranked unigrams in the positive and negative lists, we use Chi-square statistics [78] to select the most distinguishing words as the sentiment dictionary. Finally, we add negation, which usually reverses the sentiment of words (*e.g.*, "bullish" is positive while "not bullish" is negative), by searching for negation words and reversing the sentiment of affected words [79].

The resulting dictionary is then used to calculate the sentiment of a post. We label each sentence of the post to be either positive or negative, depending on whether there are more positive or negative words in the sentence. The overall post is considered positive if positive sentences outnumber negative sentences and vice versa.

For validation, we randomly sampled another 500 posts and manually labeled their sentiment. We get an overall 81.0% accuracy, with 98.8% precision and 81.3% recall for positive posts, and 13.9% precision and 75.3% recall for negative posts. Although the precision for negative posts is low, recall is high. This means that our approach is often aggressive in finding negative posts. In the context of our analysis, more false positives in detecting negative sentiment means we could over-report negative sentiment. This means our results are likely understated, and insulation from negative sentiment is likely even stronger than we report in the next section.



(a) Yahoo (b) iHub

Figure 3.7: Correlation evaluation between sentiment and stock price movements, with different time window T.

3.5 Echo Chambers

In this section we present evidence of echo chambers in the discussion boards. To do so, we bring together four sets of analyses. While our analyses are largely correlational, they do allow us to show how social phenomena in online communities can align with concretely measurable outcomes in a distinct 'social computation engine' of significant import, the stock market. The four analyses are as follows:

- *Poor User Generated Information*: sentiment in the discussion boards does not correlate with future stock price movements, and in fact stock trading strategies based on the sentiment in these boards underperform the overall market.
- *Failing to Incorporate New Information*: sentiment in the discussion boards does not correlate with recent past stock price movements.
- *Resistance to Viewpoint Diversity*: user clustering reveals that the majority of sentiment in the discussion boards may reflect the viewpoints of a very small percent of users, and that many users resist negative viewpoints.
- *Mistimed Activity Adds Noise*: posting activity correlates positively with future stock trading volume and volatility.

3.5.1 Poor User Generated Information

A central tenet of an echo chamber is that it is likely to generate information misaligned with that from outside the echo chamber. This may be due to its insular nature, social structure of its membership, or resistance to alternate viewpoints, all attributes we address below. For now, in this section we present evidence that the output of these investment discussion boards is poor with respect to the very thing these discussion are trying to predict: future stock price movements. First, we apply our two sentiment analysis methods over board posts and correlate the resulting sentiment time series with future stock price moves. As we will see, these correlations are effectively zero. For further evidence we ran trading simulations based on investment board sentiment to show that such strategies underperform simple buy-and-hold market index strategies.

Sentiment and future stock performance not correlated

We measure the correlation between the daily sentiment score and future price change for each stock on every exchange day. Messages on non-exchange days are combined into the most recent exchange day. We skip days when there is no post for a given stock, and all analyses are restricted to stocks with posts on at least 5 days, though changing this 5 day minimum threshold yields similar results.

Sentiment score is a daily measure, calculated as the ratio of positive posts over all posts for a given stock on a given day. Higher values indicating more positive daily sentiment. A potential limitation of this metric is that for some days the score is only based on a single post when there is only one post at that day. This may make the results less robust.

Price change is a daily measure, calculated as $price(X+T)-price(X)$, where $price(X)$ is the adjusted closing price for a stock on day X, and T is a time window during which

we consider the price movement.

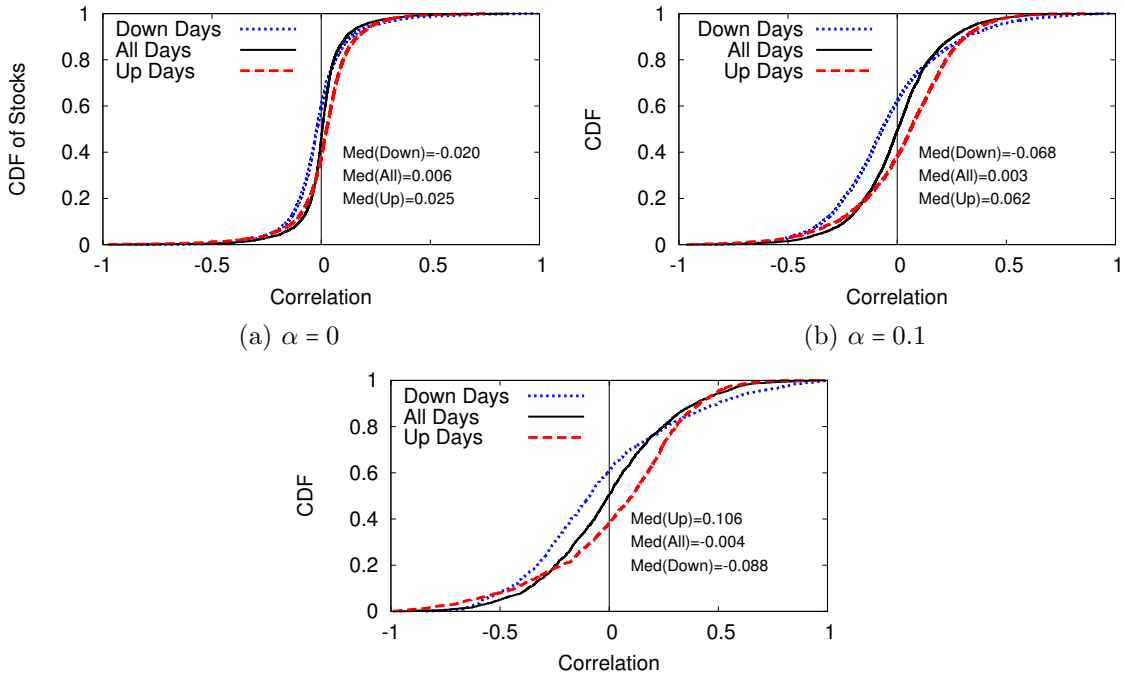
If day X has no posts for a certain stock, we skip that day. We also combine messages on non-exchange days into the most recent exchange day that has just passed. To maintain a minimum number of data points per stock, we constrain our analysis to All the analysis are done on stocks with board posts on at least 5 days[5].

Correlations are calculated by Pearson Correlation Coefficient [80], which measures the linear correlation between two variables. It is calculated as the covariance of the two variables divided by the product of their standard deviation, which produces a value in range $[-1, 1]$. A value of 1 means total positive correlation between the two variables, -1 means total negative correlation while 0 means no correlation.

We show the cumulative distribution function (CDF) of correlation coefficients with different time windows T in Figure 3.7. We make two key observations. *First*, for each dataset, the correlation is almost identical across different time windows, indicating that T has little impact on the correlation. *Second*, The majority of stocks have a correlation coefficient around 0 (the median ≤ 0.006 for Yahoo and ≤ 0.003 for iHub).

These results indicate that broadly there is almost no correlation between the sentiment around a stock in the discussion boards and future performance of that stock. However, it is possible that these summary correlations are obscuring informative signals in discussion board sentiment that is only detectable under certain conditions. For instance, perhaps the signals are weak and thus only useful in indicating very large stock price movements. Thus, we explore a variety of nuances to better understand the lack of correlation between sentiment and future stock performance. As we have consistent results for the two datasets and across different T values, for brevity we only show results for $T = 5$ day for Yahoo.

[5]We also tried thresholds of 10 days and 20 days, and get similar results.



(a) $\alpha = 0$ (b) $\alpha = 0.1$

(c) $\alpha = 0.2$

Figure 3.8: Pearson Correlation between sentiments and stock price movements for up/down/all days, with different values of price change ratio α (Yahoo).

Users are always optimistic no matter whether stock prices go up or down.

For each stock, we split its exchange days into "up days", where the stock price change is positive, and "down days" where stock price change is negative, and calculate their correlation separately. The result is shown in Figure 3.8(a). Compared to all days, which show almost no correlation, the curve of up days does have a higher correlation coefficient (median 0.025), suggesting a still very small, but non-zero positive correlation. On the other hand, the curve of down days has a median value of -0.020, indicating a very small negative correlation. While these are essentially null results, there is a small signal that when the market is bullish, investors' expectations for stock price performance are better met, while in a bearish market investors' sentiment is more irrational since they are still positive regardless how stock actually behaves.

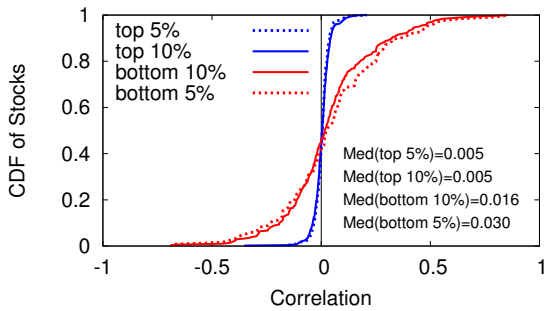


Figure 3.9: Correlation evaluation for stocks with top/bottom attention scores (Yahoo).

Users express stronger sentiment before larger stock price moves.

For each stock, we restrict our analysis to days when the relative price change ($|\frac{price(X+T)-price(X)}{price(X)}|$) exceeds a certain threshold α. Using a larger α value allows us to focus on days with larger price changes, and when $\alpha = 0$ our analysis is on all days.

The results for up days with different α values are shown in Figure 3.8. We find that the correlation tends to be more positive for larger price changes, i.e., the median correlation coefficients have larger absolute values as α increases. For down days, higher α leads to larger negative correlation. This suggests that board users are, at least to some extent, expressing stronger sentiment in advance of larger stock price moves. This increase in sentiment, however, is not any more directionally accurate and thus not useful. Again, this can be explained by the irrational positive skew towards future stock price movements among users. They are particularly bad at identifying upcoming price drops, especially when that drop is larger in magnitude.

People pay more attention to higher risk stocks.

Here we test whether the correlations are different between stocks that receive more attentions and stocks that receive less attention. As previously mentioned, we use the attention score to quantify board users' level of engagement with each stock.

Figure 3.9 shows that the correlation patterns are quite different between stocks with the top and bottom attention scores. We see stocks receiving the highest attention have almost no correlation between board sentiment and future price movement, while stocks with the lowest attention level have small but positive correlation. Inspection of high attention stocks shows that many of them are highly volatile and thus despite being inherently more unpredictable appear to attract the interest by people irrationally focused, presumably, on the high return rather than the high risk.

Trading Simulations Underperform Market

So far we have observed that there is almost no correlation between community sentiment and future stock price changes. As a further test, we developed trading strategies based on board post sentiment and then evaluated empirically whether they are profitable and can outperform the market as a whole. Generally our evaluation shows that these strategies provide limited investment benefits. Notably, for volatile and small market cap stocks, investors are very likely to lose money trading based on board sentiment.

In our sentiment-based trading strategy, a user makes investment decisions based on stock-specific board post sentiment. Specifically, each user takes a pool of money and splits it evenly on stocks whose message boards she has engaged with. For each stock, we consider two levels of ownership: a total (100%) position vs. a 50% position. If the sentiment of the previous day is positive, the user increase her position by 50% or does nothing if it is already 100%. Correspondingly, for negative sentiment the user decreases her position by 50% or does nothing if it is already 0%. For days with no posts, we do not take any trading action. We also explore other variants (0% to 100% with no increments, or increments of 33%), and observe similar conclusions. We only show the 50% increment results for brevity.

For each stock we extract the sentiment of each day by computing the ratio of positive

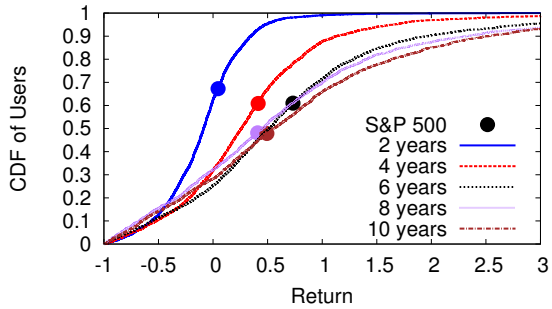


Figure 3.10: Distribution of net return using sentiment-based trading, over different simulation periods N (Yahoo).

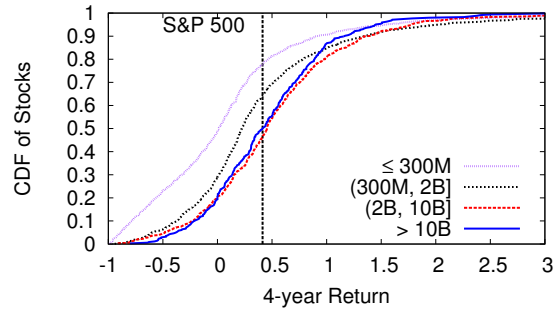


Figure 3.11: Distribution of net return by sentiment-based trading for stocks in different market cap (Yahoo).

posts on that day, and consider daily sentiment to be positive if the ratio ¿ 0.5. For days with no posts, we do not take any trading action.

We evaluate this sentiment-based trading strategy in a simulation study, based on the past N years of historical stock data[6].At the end of the simulation period we calculate the net return for each user, and compare this net return to the net return of simple S&P 500 market index.

The sentiment-based trading strategy generally underperforms the market.

Figure 3.10 shows the CDF of net returns of our trading strategy with different N values. The big dot on each line shows the return of S&P 500 index for the corresponding N value. We see that at least 50% of users underperform the market index. In particular, for shorter simulation periods ($N = 2, 4, 6$ years), 60-70% of users lose money compared to the market index. This means that in general, message sentiment is not helpful, and can even be harmful as a stock trading strategy.

The sentiment-based trading strategy performs better for larger cap stocks.

Here we explore the possibility that sentiment-based trading works better for certain types of stocks. To do so, we categorize stocks by their market cap. As mentioned above, market cap is a popular metric to compute the valuation of a company and its stock

[6]We omit stocks that do not have full price history within the past N years.

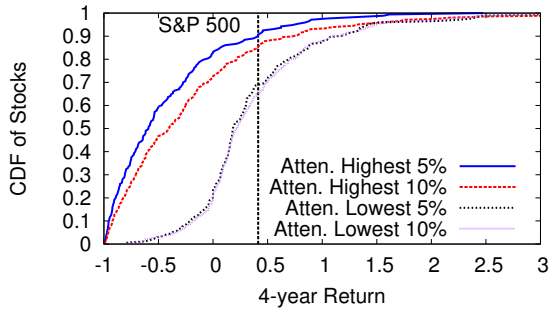


Figure 3.12: Distribution of net return by sentiment-based trading for stocks with different attention scores (Yahoo).

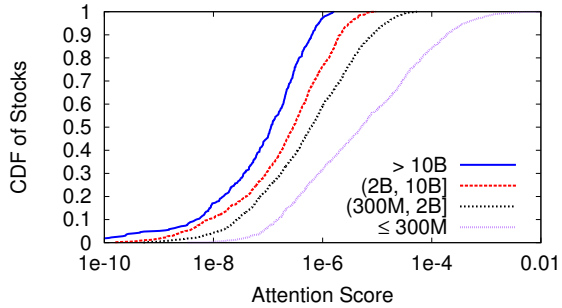


Figure 3.13: Relationship between attention score and market cap for Yahoo discussion boards.

shares. A larger market cap usually means a larger size of the company. We show the return of stocks with different levels of market cap in Figure 3.11. Generally, groups with a larger market cap have a higher ratio of stocks that beat the market return by using our sentiment-based trading strategy. For stocks with ¿$2B market cap, more than half of them have higher returns than market index. This means that for large cap stocks, trading strategies based on board post sentiment might help an investor beat the market average.

Smaller cap stocks attract more users, but perform worse.

We use our attention score as a measure of user engagement, and look at how sentiment-based trading performs on stocks with different attention scores. Results in Figure 3.12 displays different patterns for stocks with the lowest and highest attention scores. The group with the lowest attention scores has a higher ratio of stocks that beat the market return. To further investigate this, we look at how the attention score is related to market cap in Figure 3.13. The smaller the market cap, the higher the attention score, which means that small cap stocks actually attract more discussion. This suggests that stocks with the most attention actually produce the most misleading guidance for trading.

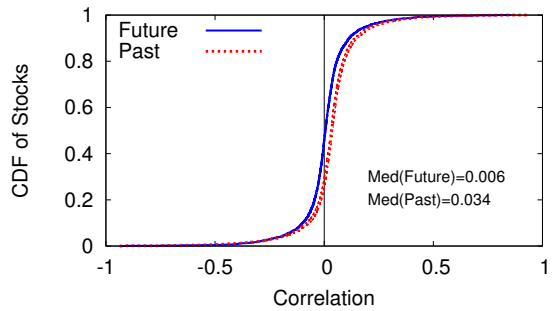


Figure 3.14: Pearson Correlation sentiment vs. future/past stock performance ("Future" vs. "Past")

3.5.2 Failing to Incorporate New Information

In the previous section we showed that these investment boards produce information that in fact is non-informative and generally will hurt investors. Here we investigate the possibility that these boards are insulated from outside information. Such insulation is both an attribute of echo chambers and a possible explanation for why these discussion board communities produce bad information. In the context of investment discussion boards, such isolation would be reflected in the inability of the community members to incorporate important new information about stocks. To test this, we examine whether the sentiment of board posts reflects changes in stock prices of the recent past. A lack of correlation with subsequent board sentiment implies that board discussion is not adequately capturing these changes. Also, given that the in the previous section we documented the lack of forward looking correlation between board sentiment and stock price movement, a lack of backward-looking correlation rules out the explanation that board sentiment simply reflects commentary on past events.

To address this, we consider past price change ($price(X) - price(X - T)$) instead of future price change ($price(X + T) - price(X)$). We denote this new correlation as "Past" and the original as "Future".

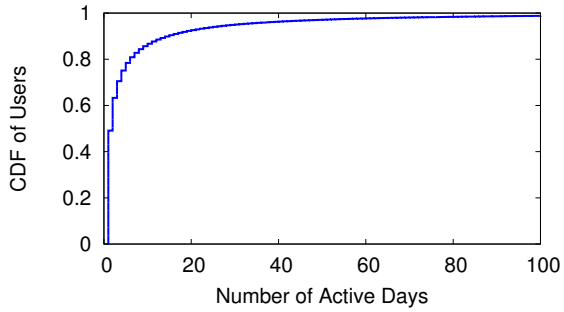


Figure 3.15: Distribution of active days over users in Yahoo message boards.

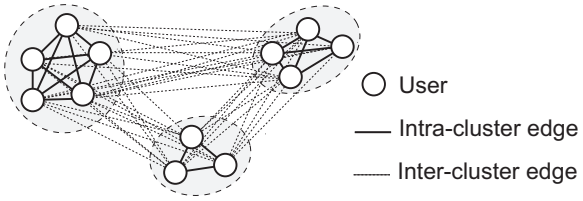


Figure 3.16: Clustering using similar graphs.

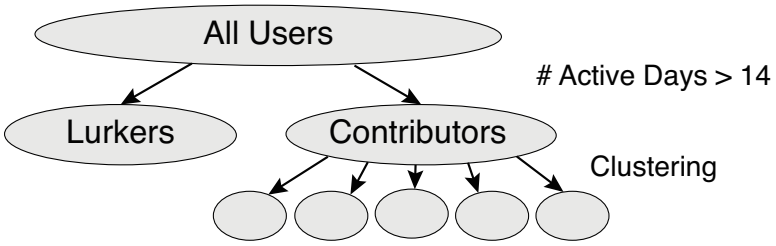


Figure 3.17: User groups extracted by our algorithm.

Figure 3.14[7] shows very similar curves for "Future" (blue line) and "Past" (red dashed line) except that "Past" has a slightly more positive correlation (with a median correlation of 0.034). Thus, while people's sentiment might reflect past stock movement marginally better than as a forecast for future movement, they still have very little correlation with stock performance, either past or future.

3.5.3 Resistance to Viewpoint Diversity

A third characteristic of an echo chamber is that only a few viewpoints are represented and that the group generally is resistant to opinion change. This is particularly damaging to the ability to make successfully predictive commentary about the stock market given that the market is an aggregation of many thousands of people. The isolated opinions of

[7]We also compare results between different strategies for up/down days and attention scores, and observe similar trends.

a few are unlike to align with the average of many. Here we identify the *roles* of board members, which capture commonalities in behavior patterns [81]. Examples of roles repeatedly mentioned in the literature include *lurkers* [82], *leaders* [83] and *chatters* [84]. Our results show that the bulk of sentiment is likely to be produced by a very small minority, and importantly that the majority of users that react to new information are highly resistant to negative sentiment.

We start by filtering out *lurkers*, who are inactive users that contribute very little content. We then do clustering on the remaining users, whom we call *contributors*. We adopt an unsupervised learning approach called *similarity graph* [85] to identify different user groups. We then study the flow of messages between these user groups and how it affects overall sentiment. In the following, we report our findings on Yahoo message boards, omitting similar results for iHub due to space limitations.

Identifying Lurkers

As reported in many online community studies [82], the bulk of users are inactive, often referred to as lurkers. As contributors are the focus of this analysis, to exclude lurkers we set a threshold on *active days* to quantify the activity level of a user, which is the total count of days that she posts messages. We show the CDF of active days per user in Figure 3.15, and observe a highly skewed distribution. We use a threshold of 2 weeks to select contributors[8], and find that these 10% of users contribute 74.5% of total posts.

Clustering via Similarity Graph

To identify behavior groups among contributors we apply the aforementioned *similarity graph* to find natural clusters among users. In a similarity graph, each user is represented by a node, and every pair of users are connected by a weighted edge, which

[8]We also used other thresholds such as 1 month, and omit similar results here.

represents the distance between them. User distance is calculated by comparing a set of behavioral features between the pair of users. Thus in the similarity graph, similar users are connected by short-distance edges. After building the similarity graph, we can extract natural user clusters by applying a clustering algorithm on the similarity graph. Figure 3.16 shows an example similarity graph.

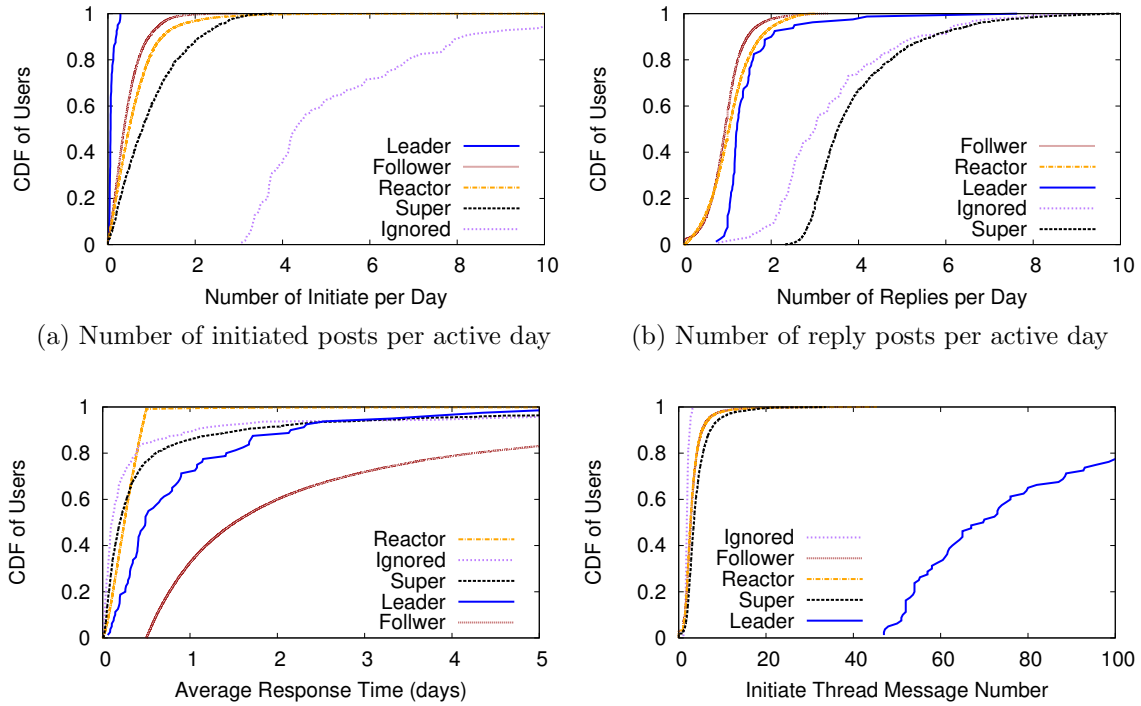
The key step in building a similarity graph is to choose appropriate features that are representative in user behaviors. We calculate the following 3 types of features to cover the most important aspects:

- *Posting behaviors*: the count of posts a user initiated and replied to per active day respectively.
- *Interaction with others*: the time difference between a user's post and the post she replies to, *i.e.*, how fast she responds to others.
- *Influence level*: the average length (*i.e.*, count of total posts in a thread) of message threads initiated by a user, where a larger value means she leads longer discussions and thus is more influential.

After calculating these features for each user, we apply Euclidean distance as the distance metric for edge weight in the similarity graph. We identify different user behavior groups by applying Divisive Hierarchical Clustering [86]. Divisive Hierarchical Clustering begins with all users in a single cluster, and continuously splits clusters. We terminate the splitting process when a further split does not improve the clustering quality, which is measured by *modularity* [87], a popular metric that compares the distance of edges inside and outside clusters.

User Behavior Groups

By applying the above clustering algorithm on contributors, we obtain 5 main clusters, with 0.03% users left as outliers, as shown in Figure 3.17. To interpret the roles of these



(a) Number of initiated posts per active day

(b) Number of reply posts per active day

(c) Average response time

(d) Length of discussion threads initiated

Figure 3.18: Feature distribution in different user groups.

user groups, we measure the distribution of each feature among different clusters, shown in Figure 3.5.3.

- *Leaders (0.09%).* These are generally discussion leaders in message threads. They do not initiate or reply to messages frequently, but when they initiate a message thread, they attract many replies (*i.e.* the blue curve in Figure 3.5.3(d)).
- *Super users (1.66%).* These users are active in both initiating posts and replying to others (the black curve in Figure 3.5.3(a) and Figure 3.5.3(b)).
- *Ignored (0.14%).* These have extremely high frequency of initiating posts (purple line in Figure 3.5.3(a)), but get very few replies (Figure 3.5.3(d)). They tend to participate in more message boards, and write short messages (*e.g.*, "START BUYING!! GO CGO!"). They are likely spam accounts trying to draw attention to certain stocks, and

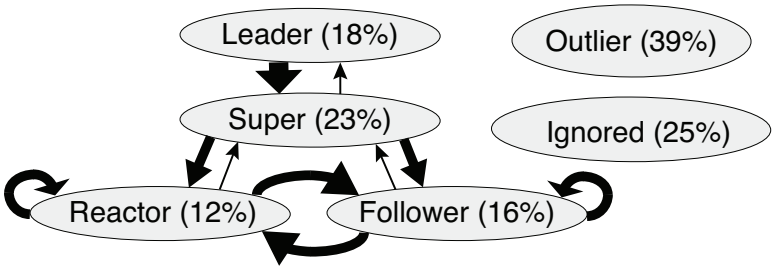


Figure 3.19: Flow of messages with negative sentiment.

are generally ignored by others.

- *Followers (61.8%)* and *Reactors (36.3%)*. These two user clusters make up the bulk of active users. What separates the two groups is the interaction pattern with others. Followers have the longest average time to respond (brown curve in Figure 3.5.3(c)), *i.e.* they tend to reply very late to posted messages. In contrast, Reactors respond to messages with the shortest average delay (orange curve in Figure 3.5.3(c)), indicating that they actively participate in discussions.

Sentiment and Message Flow

Having identified natural roles in message boards, we want to understand how they affect the flow of messages and overall sentiment. To do so, we look at our messages from the Yahoo message board dataset that come with embedded sentiment (either bullish or bearish). Note that this provides us with a ground truth dataset of sentiment in these messages about the stocks to which they refer.

We illustrate the possible directions of message flow between different groups in Figure 3.19, and note the ratio of messages authored by each group that have negative (bearish) sentiment. There are a few things worth noting. First, messages from the small "leader" and "super" groups have more negative content than their followers. More specifically, the fastest reacting "reactor" group registers a significant drop in negative sentiment from near 20% to only 12%, meaning there is a significant dilution or reversal

of negative opinion within the highly active group. It's also worth noting that the most negative sentiment is generated by those in outlier clusters, users who do not fit into the normal communication patterns of the message boards.

Finally, we find a higher level of animosity or antagonism in responses to those who post negative messages. To do so, we randomly select 200 of our ground truth sample of threads from Yahoo, and manually label all followup messages (~650 posts) as "antagonistic" or "neutral/friendly". From messages marked "antagonistic", we build a dictionary of the strongest negative keywords that capture animosity, including: `attack, blame, blind, bad, bet, blowhard, fool, failure, dad (I'm your daddy), disgruntled, ignorant, incorrect, homework (do your homework), misunderstand, newbie, ugly, sick, suck, scam, stupid, piddling, peabrained, pitiful, poor, wary, worst`.

We use this "animosity" dictionary to capture the level of personal animosity in responses to both negative and positive posts. We compare our 200 threads starting with negative sentiment against 200 threads with positive sentiment, and look at the frequency of occurrence of our dictionary words in the responses. The differences are dramatic. When we count the frequency with which our negative words occur, they account for 0.25% of all words in responses to positive posts, but 0.7% of all words in responses to negative posts. Similarly, when we look at the portion of responses that contain at least one word from our animosity dictionary, 7.8% of responses to positive posts match, compared to 19.2% for responses to negative posts. These results highlight the antogonistic response from the community to negative sentiment posts in message boards, and may help to explain why there is so little room for dissenting opinions in these forums.

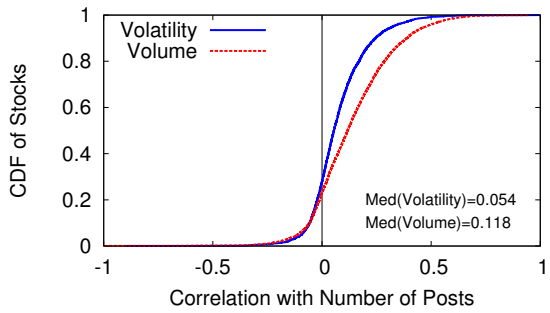


Figure 3.20: Correlation between number of posts with trading volume/volatility (Yahoo).

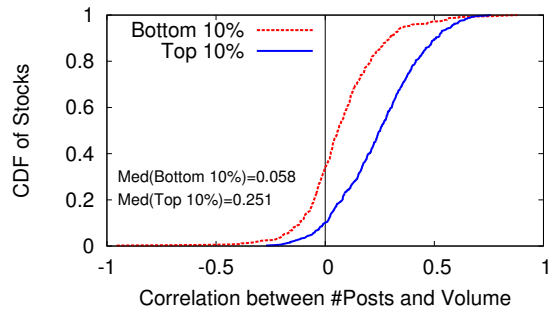


Figure 3.21: Correlation between number of posts and volume for stocks with top/bottom attention scores (Yahoo).

3.5.4 Mistimed Activity Adds Noise

We showed that investment board discussions neither captured important information from the recent past nor generated information of value going forward. Further, the bulk of the discussion is initiated by a small subset of users and many community members actively resist negative sentiment. Now we examine possible effects of users taking action based on this poor information. To be clear, we do not actually know what, if any, action users took. Instead we present evidence of the most likely outcome of people acting en mass on poor information: increased activity and variance in the market. Specifically, we show that board posting activity correlates with future trading volumes and volatility.

The trading volume of a stock on a particular day is the number of shares exchanged during that day. Volatility of a stock describes the extent to which the price changes or fluctuates. To calculate volatility, we use the standard deviation of price change during a fixed-length time range. To describe user activity, we measure the *number of posts*, which is simply how many messages appear on the discussion board.

We again use the Pearson correlation coefficient to describe the relationship between user behavior and market status. As volatility calculation requires a time range, for each stock we correlate the number of posts from day $X - W + 1$ to day X with volume or

volatility from day $X + 1$ to day $X + W$. The W-day window sweeps through the whole history of each stock. We use $W = 5$ to report our results, but other window lengths produce similar results.

Figure 3.20 shows positive correlations between number of posts and volume (median r = .11, mean r = .14) and volatility (median r = .05, mean = .08). Furthermore, when we break out stocks by their attention score, we find that the correlation is stronger in higher-attention stocks (e.g., the median correlation between post activity and volume is .25), as shown in Figure 3.21. These results suggest that user activity in these investment boards are somewhat correlated with market activity (volume and volatility), indicating a possible source of noise and inefficiency in the stock market.

3.6 Discussion

The preceding results provide evidence of the ineffectiveness of two extremely large online communities. These investment boards have the near-singular goal of helping people discuss and make sense of the stock market, and yet sentiments expressed in board discussions fail to correlate with either past or future stock price movements. If nothing else, one would imagine these boards would be vehicles for people to vent about falling stock prices, but we simply do not see that. Instead, we find that discussions are often initiated by a very small number of community members, and discussions that fail to conform to an overly positive outlook are often rejected by the community.

Methodologically, a unique feature of investment discussion boards is that they can be aligned with extremely concrete outcomes in a related but distinct environment. Here, we found non-trivial correlations between investment board activity and near-future noise in the stock market. As mentioned, we are clear that we have no way of knowing what market actions our investment board members actually took. Circumstantially, the size of

the boards in terms of numbers of people is large and our data samples cover substantial lengths of time and numbers of stocks, so it is not inconceivable that these boards are a source of noise in the market.

The financial market is a social computation engine, and such inefficiencies in this engine can have real world consequences. Even a small number of noise traders may adversely affect markets informationally and allocationally with significant price impact, by changing the prices of assets in extreme states of the economy [60]. We can easily imagine that such extreme states and price moves are much more likely in a heightened emotional state, when investors are more likely to seek conformity and herd. Thus it can create a feedback loop that is impervious to actual information. By increasing the volatility (risk), it subsequently increases the rate of return (reward) or the cost of capital any asset must offer. Based on our results where the most negative impacts are felt on small stocks, these echo chambers might have significant impact on the viability of small firms in the real economy by increasing their financing cost.

Given the fairly clear negative effects of these boards, both individually and for the market as a whole, why would people engage or pay any attention at all to these discussions? The picture painted of the bulk of active members is one of irrational investors stubbornly positive about long shot bets. These bets appear to reflect a skewed risk/reward assessment, where members are overly focused on 'big payoff' small cap stocks that in reality are also more likely to yield big losses. In itself this is damaging to these people's personal investments, as the market simply is more balanced. The stock market is extraordinarily complex. Thus it is natural that investors would look to others for guidance, akin to a social learning process. However, as we have shown here, this community of other investors is a poor group to model.

Ideally, these discussion communities are facilitating efficiency rather than noise. To leverage rather than waste the considerable amount of human attention and information

sharing that investment discussion communities support, we suggest designing for three online community characteristics in particular. First and foremost, support diversity of opinion, which almost by definition will increase market efficiency. While there is no direct evidence correlating stronger opinion diversity with outperformance, lack of diversity in this case is like extracting "wisdom" from a biased crowd. Second, ensure that new information can be easily incorporated. This could potentially be done through technological intervention. For instance, news and financial data for each company could be aggregated and kept current within the discussion forum. Finally, as with any task, feedback is critical. In this case, performance based on sentiment can be tallied and reflected back for both individuals and the community as a whole.

3.7 Conclusion

Our work studies the presence of echo chambers in investment discussion boards, and their potential contribution to noise trading behavior by participating users. Using large datasets drawn from Yahoo and iHub messages boards, we show that discussion board activity and post sentiment show negligible correlation with future market returns, and trading simulations based on board sentiment are poor investment strategies. We examine the social structure of communities inside discussion boards, and observe that the large majority of sentiment is led by a small core of users. Our results suggest that discussion boards might play a significant role in adding to noise trading by individual investors. We hope that by shedding light on the quality of these discussion boards, we encourage the adoption of forums more encouraging of divergent opinions and open exchange of ideas.

Chapter 4

Herd Behavior in Stock Markets

4.1 Introduction

Stock markets are often viewed as opaque entities that deny analysis and modeling. Yet at their core, they are effectively complex and heterogeneous social computation systems, where the collective decisions of individual investors and money managers interact to determine the value of publicly traded companies in real-time.

Despite the obvious importance of stock markets on individuals and world economies, little is known about factors that influence investment decisions by market participants, or the actual impact those decisions have on the broader markets. How much do external data influence investor behavior? How similarly do investors respond to external factors, and do they actually exhibit *herd behavior* hypothesized in behavioral finance [88]? Such important questions shed light on the power of group behavior in modern financial markets, and their answers can help guide the design of market policies and mechanisms in maturing stock markets.

Herd behavior refers to how individuals in a group can behave collectively without centralized direction. In financial markets, herd behavior is often cited as irrational

behavior that spreads through entire markets during periods of market extremes [9], either during "bubbles" created by greed, or market crashes driven by collective fear. A well-cited user survey found herding behavior spread through word-of-mouth between institutional investors [43]. More recently, empirical studies have used the metric of dispersion to assert that *market-level* herd behavior does not exist in modern markets [89, 90].

In this paper, we study the phenomenon of herd behavior at a finer granularity, *i.e.* collective behavior at the level of individual stocks instead of an entire market. Our hypothesis is that in the absence of sufficient high-qualified financial data (common in less mature markets like Chinese stock markets [91]), individual investors must make trading decisions relying on algorithmic tools such as "technical analysis," pattern matching techniques that predict future movement based solely on historical data. Additionally, information starved investors will be more likely to turn to strategies shared through social channels, further spreading the use of popular technical tools that together serve as coordination mechanisms between market participants following the same stock. This coordinated herd behavior is likely to move the price of individual stocks, and produce the same price movement the herd are expecting, *i.e.*, a self-fulfilling prophecy.

We seek to validate our hypothesis by studying and contrasting investor behavior in the Chinese and US stock markets. The Chinese stock market is a young market (25 years old), where individual investors account for more than 80% of market value and 85% of all transactions [92, 93]. In contrast, the US markets have over a century of history, and less than 30% of market value and 2% of trades (*e.g.* on the NYSE) are attributed to individual investors [94]. Given the maturity of the US market, investors are more likely to employ fundamental analysis strategies that use a variety of valuation models and company-specific financial data (*e.g.*, quarterly earnings reports) to guide investment decisions.

We first deploy a user survey, and find that while *few* US-based investors rely on technical analysis for investment decisions, *all* China-based investors in our study used technical analysis on a regular basis. A simple look at US markets fund distributions shows that technical analysis does not play a significant role for most US-based funds. This means US markets can serve as a natural baseline to help detect possible herd behavior (in Chinese markets) on specific stocks, by observing the resulting increased predictability of stock prices that match specific technical patterns.

To explore whether such heavy reliance on technical analysis in Chinese markets has actually generated herd behavior, we analyze over two decades of stock data in US and Chinese markets, and find surprisingly strong support for our hypothesis. We find that technical analysis tools are significantly more accurate, when applied to stocks in the Chinese market than those in the US market. The most popular tools (as reported by our survey results) produce the largest gaps in predictive power between US and Chinese markets, consistent with the impact of herd behavior on individual stocks observed in Chinese markets.

As a further validation that herd behavior has made Chinese markets more predictable and thus *inefficient*, we propose a simple stock trading strategy designed *solely* on identifying and benefiting from near-term herd behavior on popular technical indicators. In the context of the "efficient market hypothesis (EMH)" [67] that asset prices fully reflect all available information, technical analysis should produce no net gain in an "efficient market." Our simulation on 8+ years of historical data in both markets show that our strategy fails badly with either zero gain or net losses in the US markets. However, it significantly outperforms the market index over multiple years in the Chinese markets, providing strong evidence that Chinese markets are still inefficient.

4.2 Background and Datasets

In this section, we briefly introduce background knowledge of stock markets, basic terms in stock analysis strategies, and datasets used in our study.

4.2.1 Stock Market Basics

Stock markets offer a way for companies to raise capital from public investors in exchange for partial ownership of the company. To maximize profit from trading, investors have developed different strategies to predict future stock price movements, which can be generally classified into two categories: Fundamental Analysis and Technical Analysis.

Fundamental Analysis (FA). FA is an umbrella term for all strategies that use financial data or news to derive the intrinsic value of a company. This includes a variety of valuation models reliant on financial statements (like balance sheets, income statements), recent company news, and general world news. The derived value predicts the "intrinsic" value of a company, the assumption being that its stock valuation will correspond with this intrinsic value given enough time.

Technical Analysis (TA). TA makes predictions of future price movements using only a stock's historical data, primarily price and volume. Proponents of TA believe that, regardless of a company's health and any future events, price patterns repeat themselves. It uses statistics or chart patterns to make price predictions, which can be directly turned into trading strategies.

We define several key terms used in our analysis below.

- *Price Trend.* A general direction over time in which the price of a stock is headed: uptrend (rising), downtrend (falling) and sideways (flat). Trends can vary in length: *long term* over years, *medium term* over months, and *short term* over days to weeks.

- *Reversal.* Reversal is a change in the direction of a price trend: either from an uptrend to a downtrend, or vice versa.
- *Overbought/Oversold.* Overbought (oversold) refers to the condition where too many people have bought (sold) a stock, driving its price higher (lower) than its true value or its moving averages. The price is expected to correct itself (or revert to the mean), and fall (rise) in the near future. When a stock is overbought or oversold, traders often expect a reversal in the short term.
- *Buying/Selling Pressure.* Buying (selling) pressure is created when many people are trying to buy (sell) stocks. This causes an increase in demand (availability) for shares of the stock, which in turn drives a stock's price up (down).
- *Momentum.* Momentum measures the rate of change in a stock's price, *i.e.*, the first derivative of a stock's price.
- *Adjusted Price.* A stock's share price adjusted for stock splits and dividend payments.
- *Money Flow.* Money flow is a measurement of the approximate dollar value of one day's trading.

Formally, technical indicators are mathematical formulas based primarily on price and volume [95], and generate *signals* for trading. One popular example is the simple moving average (SMA) which smooths the stock price history by averaging prices over a multi-day window, using crossovers as signals. A bullish, or positive, crossover signal occurs when a short-term SMA crosses above a long-term SMA. This is known as a "golden cross." The reverse signal is a "death cross," which happens when a short-term SMA drops below a long-term SMA, and predicts continued drop in price in the near future. Fig 4.1 shows a death cross in the price history of Yahoo Inc. (YHOO) over 9 months (January to September 2015), which correctly predicted the continued downward movement in the share price.

Figure 4.1: A technical chart of Yahoo stock (YHOO) over 9 months (Jan. to Sep. 2015), showing a death cross on May 8th.

4.2.2 Datasets

We use the Yahoo Finance open API[1] to crawl historical stock prices for multiple stock markets. For each stock, we obtain its *daily* open and close prices, volume, and the highest and lowest price of the day up to the end of our data gathering phase, *i.e.* Oct. 7th, 2015. To remove the impact of artifacts like stock splits and dividends, we use the adjusted closing price from Yahoo Finance and compute the adjusted daily volume to match.

We summarize the final datasets used in our analysis in Table 4.1. For both US and China, we focus on the two largest stock exchanges: New York Exchange (NYSE) and NASDAQ for the US market, and Shanghai Stock Exchange (Shanghai) and Shenzhen Stock Exchange (Shenzhen) for the Chinese market. To match the Chinese dataset and control variance, we limit the U.S. market data to only include the last 25 years, starting from January 1st, 1991 to October 7th, 2015. Our dataset includes all stocks in each exchange within this period, with the exception of: 1) preferred or special share variants, 2) new stocks that started trading after June 30th, 2015, and 3) de-listed stocks.

[1] http://finance.yahoo.com/

Country	Exchange	Total Stocks	Covered Stocks	Data Since
U.S.	NYSE	3278	2854 (87%)	01/02/1962
	NASDAQ	2964	2846 (96%)	01/02/1970
China	Shanghai	1114	1113 (100%)	12/19/1990
	Shenzhen	1729	1711 (99%)	01/02/1991

Table 4.1: Summary of stock coverage for both U.S. and Chinese markets. Our final stock datasets include all the covered stocks over 25 years, between Jan 1st, 1991 to Oct. 7th, 2015.

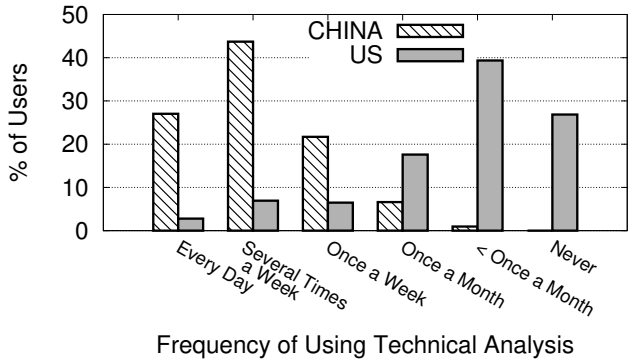


Figure 4.2: Frequency of applying Technical Analysis (TA) tools, China versus US.

4.3 User Survey: Personal Investment Behavior

We begin by trying to identify factors that help determine how investors choose their trading strategies, and compare results for investors across Chinese and US markets. We deploy a user survey using SurveyMonkey in the US, and a translated version on its equivalent counterpart in China (Wenjuanxing)[2].

Survey setup We launched our user study in the U.S. and Chinese markets simultaneously in January 2016. We deployed the survey using popular online survey services: SurveyMonkey[3] in US and Wenjuanxing[4] in China. Both services provide large user pools from which we can collect responses and target certain demographics of users. Since Sur-

[2]Our study obtained institutional IRB approval.
[3]http://www.surveymonkey.com/
[4]http://www.sojump.com/

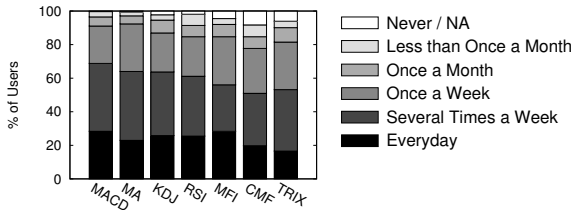


Figure 4.3: Usage frequency for different indicators among Chinese investors.

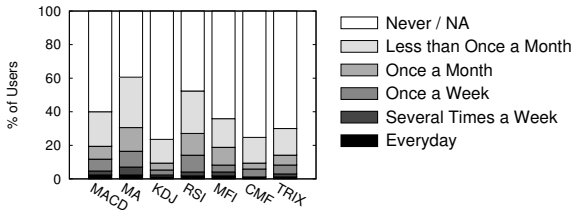


Figure 4.4: Usage frequency for different indicators among US investors.

veyMonkey does not provide a function to target stock investors, we instead target users with general investment experience. In total, we collected 314 responses from Wenjuanxing and 302 from SurveyMonkey. After filtering out users who did not invest in stocks, we totaled 314 valid responses from Wenjuanxing and 216 from SurveyMonkey.

For questions we asked, first, we determine each user's investing experience and risk tolerance. Second, to determine users' investment strategies, we ask the frequency they consult technical analysis and fundamental analysis, and different technical indicators if applicable. Third, to understand why people choose certain investment strategy, we ask them to provide the primary reason(s) for selecting a specific strategy, and ask how their strategy would change given more reliable financial information. We then ask users what additional resources would help them make better investment decisions. Finally, we collect demographic information (*e.g.* age, income, gender). We attach our survey questionnaire in Appendix A.1.

Demographics. In terms of respondent demographics, we observe very similar distributions for gender (male slightly more than female) and levels of risk tolerance between both markets. 70% of users have moderate risk tolerance: they are willing to accept the occasional loss as long as their money is in sound investments. Of the rest users, half have low risk tolerance and half have high tolerance.

The distributions of age and investment experience differ between the two markets. Chinese investors tend to be younger and much less experienced. The median age of

Chinese respondents are between 20 to 30, while for US respondents the median age is between 50 to 60. In terms of Investment experience, the median for Chinese respondents and US respondents are ¡10 years and 10-20 years, respectively. This is as expected, since the Chinese stock market has only been around for 25 years while the US market is more than a century old.

Group Behavior in Investment Decisions. We asked how often investors apply TA (technical analysis) in their investment decisions, and plot the distribution of usage frequency for both markets in Fig 4.2. The difference is significant. All China-based investors in our survey used TA, while 27% of US responders never use it. Furthermore, more than 90% of Chinese investors use TA once a month or more, compared to only 16% of US investors. In fact, our survey shows that US investors prefer a much more passive approach to investing. 80% of them consult either TA or FA (fundamental analysis) less than once a week, in contrast to only 10% of Chinese investors. 14 US respondents (~6%) explained that their financial advisors make actual investment decisions. In particular, one user mentioned: "I try to diversify, and based on overall history over at least 10 years. I check these frequently but only change allocations once every two years on average."

Of the responders who said they use TA, we asked them which technical indicators they rely on the most. Fig 4.3 and Fig 4.4 plot the usage frequency of the 7 key technical indicators for Chinese and US investors, respectively. We make two observations. *First*, Chinese investors rely heavily on these indicators. Our data shows that 68% of our responders use *all of them* at least once a week, and 78% use all of them once a month. In contrast, US investors only use them occasionally and are more diverse in terms of the choice of the indicators. *Second*, while all 7 indicators are popular among Chinese investors, MACD, MA and KDJ as the top 3 indicators. For US, the top 3 choices are MACD, MA and RSI.

Possible Causes for Group Behavior. Our results show that the majority of Chinese investors use the same inputs in their decision making, *i.e.* TA with a small set of popular "technical indicators." This contrasts with the low utilization of TA among US investors, and the lack of reliance on the key technical indicators.

But what are factors that led to this surprisingly consistent behavior in the Chinese investment community? In our survey, we asked users to indicate the key reasons triggered them to use TA tools. For both Chinese and US responders, the most common reason is the inclusion of these tools in stock investment softwares. Beyond that, US investors report that they enjoy the simplicity of TA while Chinese investors are heavily influenced by their social circles and social media sources, *e.g.* online blogs, forums and social networks. This is consistent with the anecdotal evidence that in China, personal investments are viewed as both a social activity as well as an indicator of their personal value [96].

We also observe a stronger self-reliance on personal decision making in Chinese investors compared to their US counterparts. When asked what they would do if reliable data about companies and the market were more available, 65.6% of Chinese investors said they would not change their investment strategy, compared to only 40.7% of US investors. When asked what additional resources will help them with investment, 66.2% of Chinese investors prefer subscriptions to investment/financial newspapers or journals, while US investors chose personal investment consultants (40.3%) and additional investment/stock market news (36.6%). This highlights the difference in the amount of financial data available to individual investors between the two countries, and confirms the fact that financial data is less available in China due to lack of regulation on corporate financial transparency [97].

Summary. Our user study on both Chinese and US investors show dramatically different investor strategies. US investors tend to be more passive and reliant on external

advisors. Those who use technical analysis consider a wide range of tools and indicators. In contrast, Chinese investors are more self-reliant, but have less information available to them. All of our Chinese participants actively use technical analysis and rely heavily on the same small set of popular indicators, possibly due to heavy influences from social circles and online media.

4.4 Macro Impact of Group Behavior

Our user study revealed that Chinese investors relied heavily on technical analysis techniques, much more than all other information sources combined (which we refer to as fundamental analysis). The fact that they all used a small set of common technical indicators included within stock trading software suggests that the majority of Chinese individual investors follow recommendations generated by the same algorithms. Since the Chinese market is largely dominated by individual investors, this herd behavior is therefore likely to generate a visible, non-trivial impact on the prices of individual stocks [98, 99]. This is consistent with the theory that technical analysis tools succeed due to a "self-fulfilling prophecy" effect, *i.e.* predictions of stock movement come true because enough investors follow the signal as a herd, thereby generating the predicted price movement through their own actions.

The natural question is, *has the collective reliance of Chinese investors on a few technical analysis indicators generated herd behavior*? And if so, how strong is the impact of this herding behavior, and is it detectable across the entire market? In this section, we describe our experimental methodology to try to answer these questions, by analyzing the "accuracy" of technical analysis tools in the Chinese stock market, using the US stock market as a baseline for comparison.

4.4.1 Detecting Herd Behavior in Stocks

A self-fulfilling prophecy exists where belief in a system or process causes the system to become true. It is commonly believed that there exist self-fulfilling aspects to technical analysis [100, 101]. There have been quantitative studies on the subject, with positive results in support of this assertion [102, 103]. This assertion argues that the key to a technical indicator's usefulness lies in its popularity: the easier it is understood, and the more broadly it is applied, the more accurate its predictions become.

If true, herd behavior based on a small set of technical indicators will have significant implications on the dynamic processes controlling prices in the Chinese stock market. We believe the Chinese and US stock markets provide an ideal context to test the validity of this hypothesis. Results from our user study show that all individual investors in China utilize technical analysis, and rely heavily on a small set of popular technical indicators. In absolute terms, Chinese investors trade frequently using technical analysis (Fig 4.2). And since individual investors control roughly 81% of the Shanghai exchange by value and account for 85% of all trades [93, 92], herd behaviors in this user population are likely to drive significant movements in the price of many stocks.

The US stock market provides an ideal baseline or control group, where individual investors own less than 30% of the market and account for less than 2% of trading by volume [94]. Instead, much of the market is controlled by large conservative pension funds and active hedge funds. A close look at the largest 50 hedge funds by asset value shows that together they control 2.6 Trillion USD under management[5]. In descriptions of management methods reported by these 50 funds, only 3 mention some type of "quantitative," "systematic," or "multi-strategy" approach, often used codewords for trading based on technical analysis. All others cite fundamental research as their key investment

[5]http://www.octafinance.com/hedge-funds/top-hedge-funds

style. Combining this with the low use of technical trading by individual investors, it is clear that technical trading plays a relatively small role in US markets.

Quantifying the Impact of Herd Behavior. Our challenge is to detect the impact of herd behavior in investments from looking at aggregate stock price data. If millions of Chinese individual investors make stock trades based on a small group of technical indicators provided by popular stock trading software, their trades should be able to move the majority of individual stocks in the market. Proving causality is impossible without an active experiment involving large groups of investors. Instead, we must search for support for our hypothesis by analyzing the historical prices of individual stocks, and quantifying how well they follow predictions of technical indicators.

If technical analysis indicators had no correlation with fundamental value of a stock, one might expect that its predictions are "right" around 50% of the time. In our study, we apply our technique to stocks in US markets as a control group, because both US-based funds and individual investors seem to make little use of technical analysis indicators.

In the rest of our paper, we quantify the impact of this "herding behavior" by analyzing the predictive accuracy of popular technical indicators for all stocks in key US and Chinese stock exchanges. We quantify the "predictive power" of technical indicators on each stock, by computing all potential signals generated by technical indicators from its price history, and comparing them to the stock's price movement at different points of time following the signal.

4.4.2 Analyzing Technical Indicators

We focused on a group of technical indicators that are easily understood, widely used, and have clear definitions. Combing through literature on technical analysis, we end up with 8 indicators. Among them, MACD, SMA/EMA, KDJ and RSI are the most pop-

ular. They are provided by popular stock trading software and financial websites, and mentioned frequently in stock forums and research papers. They are also the most popular indicators in responses to our user survey. Specifically, MACD is often considered the most effective and popular indicator available. SMA and EMA are the simplest indicators, and often serve as a first step towards calculation of more complicated indicators. RSI is not as widely used as the others, while KDJ is extremely popular only in China[6].

We also implemented several other indicators, including MFI, CMF and TRIX. MFI and CMF track money flowing in and out of each stock by volume. MFI is also known as volume-weighted RSI, since it uses a formula similar to RSI. Our results show that MFI and CMF closely resemble RSI and KDJ, so we omit them for brevity. Finally, TRIX is known to be functionally similar to MACD, but far less popular as an indicator in both U.S. and China. We compare TRIX to MACD to see if can quantify the impact of popularity on the accuracy of functionally similar statistical measures.

We describe these indicators in detail below:

- *MACD (Moving Average Convergence Divergence)* tracks dynamics of two different moving averages of price, subtracting the longer moving average from the shorter moving average. It is designed to both inform the current price trend and capture price momentum.
- *SMA (Simple Moving Average)* and *EMA (Exponential Moving Average)* smooth price data to detect price trends by averaging recent close prices. SMA simply takes the average while EMA weighs recent days more heavily.
- *KDJ Indicator* follows the price momentum instead of the price itself, by measuring the current close price relative to the high-low price range over a period of time. It is also used for identifying overbought/oversold levels.

[6]KDJ is often listed in Chinese articles as top popular technical indicators but rarely in articles from the US, *e.g.*, http://study.huagu.com/j/cgjq/dxjq/1307/32827.html

- *RSI (Relative Strength Index)* tracks the price momentum by comparing the magnitude of recent positive changes to recent negative changes. It can either determine overbought/oversold conditions, or identify the price trend.
- *MFI (Money Flow Index)* predicts the reliability of the current price trend and warns of potential reversals, by measuring buying and selling pressure with recent price movements and volume.
- *CMF (Chaikin Money Flow)* determines the relative buying/selling pressure by summing the money flow for a recent time period.
- *TRIX (Triple Exponential Average)* also reveals the price trend and measures the momentum, like MACD. Because it applies a moving average three times to the recent close prices, it significantly reduces noise to focus on more significant price movements.

We list our implementation details in Table 4.2, where for each indicator we include both its calculation and the rule of thumb that generates a buy or sell signal. All rules for generating signals are consistently defined in both U.S. and China. However, we note that for the RSI in China, the crossover trading rule is more commonly used than the breaking range trading rule. The opposite is found in the U.S. For this case, we implement and analyze both.

Indicators typically generate signals in one of two ways:

Breaking out of a range. The indicator has a function f that fluctuates within some range. f generates a signal when it breaks through the range either above or below.

Generating a crossover. The indicator has two functions f_1 and f_2, and generates a signal when they cross. We identify that f_1 crosses f_2 from above at day x as: for some previous time f_1 is below f_2 for at least δ while at x it is above for at least δ, where δ is a single variable to quantify how "clear" the crossover is. We calculate the distance between f_1 and f_2 at day x as $f_1(x) - f_2(x)$ for range bound functions, while

Indicator	Calculation	Trading Rule to Generate Signals	Parameter Setting	Prediction
MACD	$MACD(x) = EMA_m(close(x)) - EMA_n(close(x))$ $Signal(x) = EMA_p(MACD(x))$	Buy: MACD crosses Signal from below Sell: MACD crosses Signal from above	$(m, n, p) = (12, 26, 9)$	Short-term
SMA EMA	$SMA_m(close(x)) = \frac{1}{m}\sum_{k=x-m+1}^{x} close(k)$ $EMA_m(close(x)) = \alpha close(x) + (1-\alpha)EMA_m(close(x-1)), \alpha = 2/(m+1)$	Buy: $\text{SMA}_m(\text{EMA}_m)$ crosses $\text{SMA}_n(\text{EMA}_n)$ from below, $m < n$ Sell: $\text{SMA}_m(\text{EMA}_m)$ crosses $\text{SMA}_n(\text{EMA}_n)$ from above	$(m, n) = (50, 200)$	Mid-term
KDJ	$K(x) = EMA_n(100 \times \frac{close(x)-\min(low(x))}{\max(high(k))-\min(low(k))})$ $D(x) = EMA_n(K(x))$ $J(x) = 3K(x) - 2D(x)$ min(max) is minimal(maximal) during a period of m days from day $x-m+1$ to day x	Buy: J touches DOWN from above Sell: J touches UP from below	$(m, n) = (9, 5)$ (DOWN,UP)= $(0, 100)$	Short-term
		Buy: K crosses D from below Sell: K crosses D from above	$(m, n) = (9, 5)$	
RSI	$Gain(x) = close(x) - close(x-1)$ if $close(x) > close(x-1)$, 0 otherwise $Loss(x) = close(x-1) - close(x)$ if $close(x) < close(x-1)$, 0 otherwise $AvgGain(x) = [(m-1) \times AvgGain(x-1) + AvgGain(x)]/m$ $AvgLoss(x) = [(m-1) \times AvgLoss(x-1) + AvgLoss(x)]/m$ $RSI_m(x) = 100 - 100/(1 + AvgGain(x)/AvgLoss(x))$	Buy: RSI_m crosses RSI_n from below, $m < n$ Sell: RSI_m crosses RSI_n from above	$(m, n) = (6, 12)$	
		Buy: RSI_m touches DOWN from above Sell: RSI_m touches UP from below	$m = 14$ (DOWN,UP)= $(30, 70)$	Short-term
MFI	$RawMF(x) = Volume(x) \times (high(x) + low(x) + close(x))/3$ if close(x) ¿ close(x-1), $RawMF(x)$ considered $PositiveFlow(x)$, otherwise $NegativeFlow(x)$ $MoneyFlowRatio(x) = \sum_{k=x-m}^{x} PositiveFlow(k) / \sum_{k=x-m}^{x} NegativeFlow(k)$ $MFI(x) = 100 - 100/(1 + MoneyFlowRatio(x))$	Buy: MFI touches DOWN from above Sell: MFI touches UP from below	$m = 14$ (DOWN,UP)= $(10, 90)$	Short-term
CMF	$MoneyFlowVolume(x) = ((2close(x) - high(x) - low(x))/(high(x) - low(x))) \times volume(x)$ $CMF(x) = \sum_{k=x-m+1}^{x} MoneyFlowVolume(k) / \sum_{k=x-m+1}^{x} volume(k)$	Buy: CMF touches UP from below Sell: CMF touches DOWN from above	$m = 20$ (DOWN,UP)= $(-0.05, 0.05)$	Short-term
TRIX	$TripleEMA(x) = EMA_m(EMA_m(EMA_m(close(x))))$ $TRIX(x) = (TripleEMA(x) - TripleEMA(x-1))/TripleEMA(x-1)$ $Signal(x) = EMA_n(TRIX(x))$	Buy: TRIX crosses Signal from below Sell: TRIX crosses Signal from above	$(m, n) = (15, 9)$	Short-term

Table 4.2: Detailed calculation and signaling strategies for the 8 chosen technical indicators. Notations: all indicators are functions of days, denoted as $function(x)$ at day x. *open*, *close*, *high*, *low*, and *volume*: adjusted prices and volume. m, n, p: time windows (in days) during which an indicator is calculated. UP, DOWN: upper and lower bounds that trigger trading signals.

$\frac{f_1(x)-f_2(x)}{\max(f_1(x),f_2(x))}$ for the rest.

Where appropriate, we always adopt the default parameter settings [7] for each indicator (listed in Table 4.2). Based upon our settings, all indicators are designed to detect short-term trends, with the exception of SMA and EMA, which target medium time frames.

4.4.3 Computing Indicator Accuracy

To perform our analysis, we systematically apply each of our technical indicators to all stocks in our dataset of US and Chinese stocks. For each stock and indicator combination, we detect all possible signals in the stock's price history, and then compare the signal prediction with the actual price movement after some time window has passed. We compute the predictive accuracy I_S of technical indicator I for stock S as the portion of all signals in its price history that had positive correlation to actual price movement.

For an entire stock market, we describe the prediction power of a technical indicator

[7] Default parameter settings are from either popular financial websites like *Investopedia*, or widely-used stock investment softwares like *Yahoo Finance.*

by aggregating its prediction accuracy over each stock in the market. The result is plotted as an accuracy CDF, where each data point represents a single stock.

4.5 Experimental Analysis

In this section, we describe data analysis results that search for possible impact from herd behavior in Chinese stock markets. We compare the prediction accuracy of technical analysis between the U.S. and Chinese stock markets (using both quantitative and qualitative methods), and find that technical indicators consistently show much stronger accuracy in the Chinese market.

These observations support our hypothesis that aggregated use of technical indicators by large groups of investors has had a significant impact on market behavior. Consistent with the theory that aspects of technical analysis act as a self-fulfilling prophecy, we find that the more popular the indicators in China, the larger the gap in prediction accuracy compared to that in the US market.

4.5.1 Metrics

We compute prediction accuracy CDFs for each indicator for each stock in US and Chinese markets. To quantify the differences between accuracy CDFs, we use two statistical metrics:

KolmogorovSmirnov test (KS test). It tests the hypothesis that two sample distributions are drawn from the same distribution. *p-value* is to quantify if the test outcomes are significantly different in the two samples, where *p-value* ¡ 0.05 indicates significant difference. *KS statistic* represents the maximal distance between the two sample distributions captured by a single variable.

Difference in Areas Under the Curve (AUC). The AUC of a CDF is the accumulation of CDF values along all possible values of the variable X, which can be calculated as $1 - E[X]$. Difference of AUCs between two CDFs shows their accumulated difference.

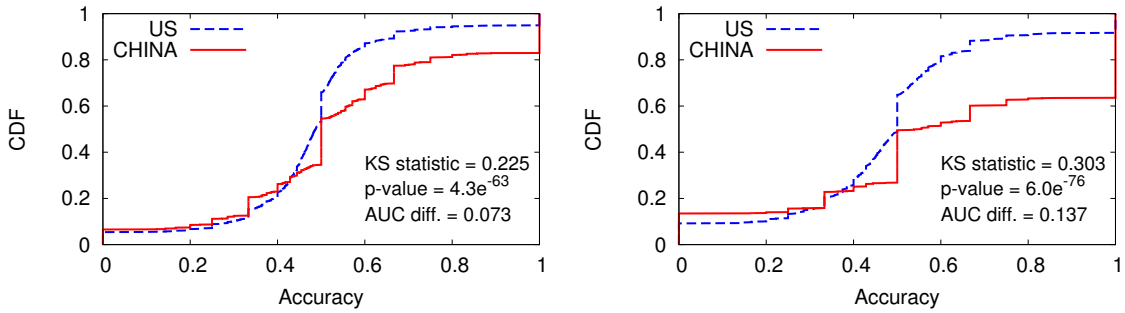
4.5.2 Performance of Technical Indicators

In this subsection, we highlight key results for our 8 chosen technical indicators, then explore individual exchanges' impact on prediction accuracy, *i.e.*, we compare stock exchanges within each of the two countries. We find that all indicators are consistently more accurate in predicting price movements in the Chinese markets than in the US markets. This higher level of prediction accuracy correlates with our findings that technical analysis is a more popular method of guiding investment decisions in China than in the US; it implies the widespread usage of technical indicators in China by the large pool of retail investors reinforces the predictions of the indicators themselves. Because our results are largely consistent, we show only highlights here due to space constraints.

For each indicator, we set the trend window parameter T as discussed in the prior section: medium-term predictions (SMA/EMA) have T = 8 months, and short-term predictions have T = 3 months. The only exception is CMF, which has a super short-term prediction, with T = 1 week. For indicators with crossover variable δ, we choose an appropriate value and leave detailed discussion later in this subsection. For RSI and KDJ (each of which has two implemented trading rules) we show results which are more statistically meaningful.

We highlight our findings below:

Indicator Performance. MACD is often considered the most effective and popular indicator in technical analysis. If our hypothesis holds, and greater reliance on technical analysis is by a large individual investor population in China, we would see a quite

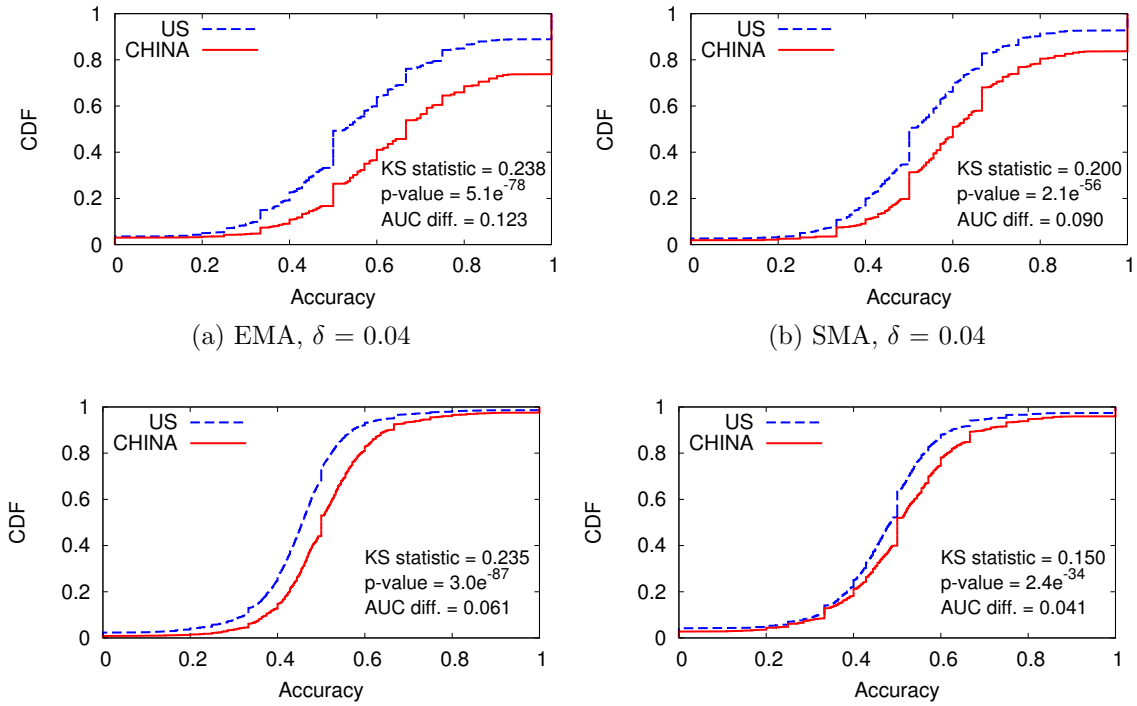


(a) $\delta = 0.3$ (b) $\delta = 0.6$

Figure 4.5: Prediction accuracy of MACD in U.S. and Chinese markets with different crossover variable δ.

significant gap in accuracy CDFs of MACD between the two countries. The results in Fig 4.5a are consistent with our hypothesis. The solid red line representing the Chinese market is located far to the right of the U.S. market (blue dashed line), showing that a far greater number of stocks in the Chinese market are highly amenable to accurate prediction using MACD signals. It's worth noting that for a surprisingly high 19% of all Chinese stocks, the MACD signals *always* correctly predict price movement when δ=0.3. Results from statistic metrics are shown inline in each figure. They also verify the significant gap quantitatively, *e.g.*, the p-value $\ll 0.05$, which strongly supports that the two CDFs are qualitatively different.

We provide results for other indicators in Fig 4.6, where we consistently observe more accurate prediction in the Chinese market than in the U.S. market. In every case, p-values are tens of orders of magnitude below any meaningful threshold. Specifically, we observe significant gaps between predictive accuracies of the two countries in EMA and SMA, which are particularly accurate in China, and able to provide better than random predictions for 70% of all stocks! These results seem to support our hypothesis, that herding behavior based on technical analysis has qualitatively changed stock pricing in the Chinese stock market, making most stock prices easier to predict.



(a) EMA, $\delta = 0.04$ (b) SMA, $\delta = 0.04$

(c) KDJ (d) RSI, $\delta = 15$

Figure 4.6: Prediction accuracy of different indicators between the U.S. and Chinese markets.

Comparing Individual Stock Exchanges. To ensure that our observations are not artifacts inherent in specific stock exchanges, we examine the two markets to quantify any performance gaps between stock exchanges in the same country, *i.e.*, NYSE vs. NASDAQ in the U.S., and Shanghai vs. Shenzhen in China. We provide the accuracy CDF of MACD for the four exchanges separately in Fig 4.7, and omit similar results for other indicators for brevity.

We make two important observations. *First*, the gaps in accuracy CDFs across countries remain significant regardless of the specific exchange. *More importantly*, accuracy CDFs for stock exchanges in the same country are very similar. This confirms that whatever factors are contributing to the accuracy results, they are specific to the country, and not individual stock exchanges.

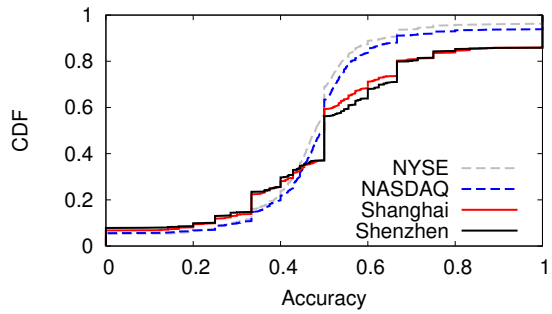


Figure 4.7: Prediction accuracy of MACD for individual exchanges in the U.S. and China. $\delta = 0.3$.

Impact of Crossover Threshold δ. We expect our observations for each indicator to remain robust across different δ values, and our results support this. While we show results for MACD, results are consistent for other indicators. We plot accuracy CDFs with two different δ values in Fig 4.5 (We tested a wide range of δ values, all showing the same qualitative results). We arrive at two critical observations. *First*, the accuracy of technical indicators is consistently higher for the Chinese market. *Second*, when δ increases the gap between the two accuracy CDFs becomes more pronounced. When δ is small, signal detection is sensitive and small fluctuations due to statistical noise can generate fake signals. For larger δ values, signals are only generated for more confirmed crossovers, the accuracy for both markets increase, and their gap grows larger, while fewer signals are generated per stock. Therefore, the goal is to choose δ values that are large enough to avoid false signals, yet small enough to generate a large sample set to draw statistically meaningful conclusions.

Correlating Indicator Popularity and Accuracy. Working with the assertion of self-fulfilling aspects in technical analysis, the popularity of an indicator should play an important role in its usefulness. The more widely an indicator is used, the more accurate its predictions should be. As a result, we expect to see a bigger gap between the two markets for the more popular indicators.

δ	KS-test (p-value)	KS statistic	AUC Difference
0.1	$3.7e^{-54}$	0.187	0.029
0.3	$4.3e^{-63}$	0.225	0.073
0.6	$6.0e^{-76}$	0.303	0.137

Table 4.3: Statistic measurements on pairs of accuracy CDFs of U.S. and Chinese markets for MACD.

We test this hypothesis by performing a comparison of relative predictive power across all indicators, and comparing that to indicator popularity in China. We note that while predictions are consistently more accurate in China, the size of the gap also correlates with the popularity of the indicator.

To quantify this correlation, we rank indicators by two criteria, their relative popularity as indicated by our user study, and the size of the accuracy gap between US and Chinese markets quantified by AUC difference. We quantify the indicators' relative popularity, by sorting them by the ratio of users who indicate that they frequently use a given indicator[8]. We list the two rankings in Table 4.4, and see a strong correlation between the popularity of an indicator and the size of its accuracy CDF gap. In other words, more popular indicators typically generate larger difference in prediction accuracy between the markets. These observations support our hypothesis that "herd behavior" by Chinese investors based on a small set of technical indicators serves to increase the accuracy of the most popular indicators.

4.6 Efficiency in Chinese Markets

Finally, we want to validate our hypothesis that herd behavior has made the Chinese stock market predictable and *inefficient*. So we backtrack historical stock prices in Chi-

[8] "Frequently" defined as using it more than once a month, but using a threshold of once a week produces similar results

Rank by relative popularity	Rank by accuracy difference
SMA/EMA	**SMA/EMA**
MACD	**MACD**
KDJ	**KDJ**
RSI	TRIX
MFI	RSI
TRIX	MFI
CMF	**CMF**

Table 4.4: Ranking of indicators by relative popularity from our survey (% of frequent usage), and accuracy difference across the two markets (AUC difference).

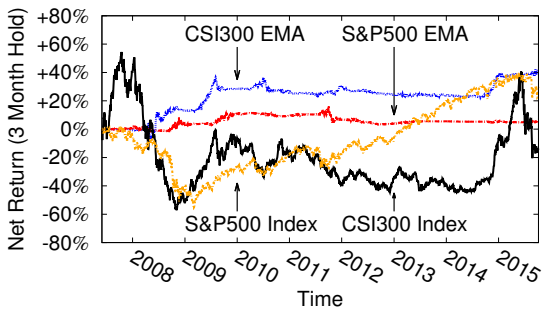


Figure 4.8: Return of simulated trading strategy using EMA signals.

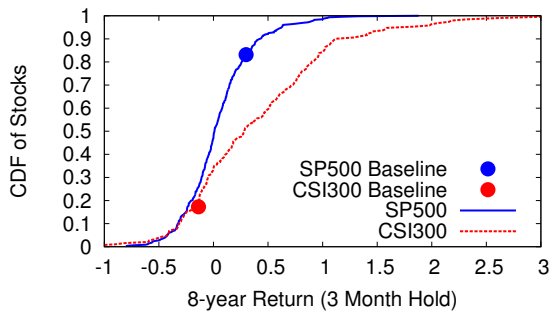


Figure 4.9: Distribution of 8-year return using EMA signals.

nese markets for a period of 8 years, and simulate simple trading strategies that rely only on signals generated by technical indicators. In an efficient market, technical analysis should produce no net benefit over random choice, and produce either zero net returns or a negative return over time. Again, we run the same simulation on both Chinese and US markets, and use the US results as the control group.

4.6.1 Technical Trading Strategy

We drive our technical trading strategy using signals generated by technical indicators. A hypothetical investor begins with an initial amount of money, and divides it evenly among a set of stocks. For a single stock, if/when a buy signal is observed, the investor buys the stock using the amount allocated to the stock, holds for some fixed time period

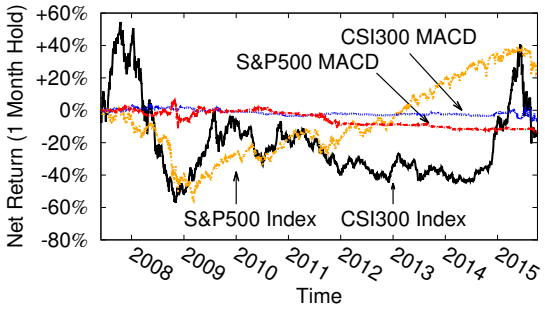


Figure 4.10: Return of simulated trading strategy using MACD signals.

T, and then sell the position. In contrast, if/when a sell signal is detected on a stock, the investor proactively "shorts" the stock using the allocated funds, and buys the shares back after time T. "Shorting" is when an investor borrows shares and sell them at the current price, with the goal of buying those same shares back at a lower price. If no positive or negative signals are found, then allocated funds for a given stock remain as cash.

We vary the value of T over different short time period (1 month or 3 months depending on the indicator), and calculate the total return as ($earned / $basic) after different time periods. To represent the US market, we split funds across stocks represented in the S&P500, which includes the 500 large companies listed on NASDAQ and NYSE. For Chinese markets, we rely on a similar index called CSI300, which captures a representative set of 300 stocks traded on the Shanghai and Shenzhen Exchanges.

4.6.2 Empirical Results

We run our simulated trading strategies between June 2007 to Oct. 2015, a period covering the global financial crisis and its recovery, during which the Chinese market was highly volatile and produced slightly negative returns. We use 3 most popular technical

indicators, EMA/SMA and MACD[9]. Given the length of our simulations, we omit stocks with < 8 year history, *i.e.* they were not publicly traded companies for the duration of our simulations.

Fig 4.8 shows the daily return over the whole simulation period for EMA, compared to market baseline. We omit the extremely similar SMA results. The results clearly show that while there is almost zero return when trading based on EMA in the US market, the same strategy produces a significant positive return (~40%) in the Chinese market. More importantly, we find that technical trading in Chinese market significantly outperforms the CSI market index (40% return vs. -13%). The results are flipped in US markets, where the technical trading strategy produces near flat returns compared to a 20+% return for the market index. We also note that this strategy is somewhat conservative, since funds allocated for many stocks remained in cash for long periods because no technical signals were observed.

To get a more detailed view, we plot the return of individual stocks using this strategy in Fig 4.9 following the end of the 8 year period. We find that more than 80% of stocks in Chinese market perform better than the market index (CSI), and only 20% stocks in US market do so against the S&P. The outperformance is extremely strong in some Chinese stocks, where individual returns went as high as 300% using only short termed technical trading.

We plot similar results for MACD in Fig 4.10. The results are consistent in that technical-based trading in Chinese markets significantly outperforms trading in US markets. But while trading on MACD produces a near-zero return in Chinese markets, it produces a notable loss in the US markets.

[9]We set EMA/SMA with T = 3 months, while MACD is a more short term indicator with T = 1 month. We varied T to confirm that these settings generate the highest returns for each indicator in both markets.

4.7 Related Work

Profitability of Technical Analysis. While our work quantifies the impact of aggregate investment behavior on price across different stock markets, many studies have studied the profitability of technical analysis in single stock market [104, 105, 106, 107].

Comparison of Different Markets. Later work tried to compare different markets [108, 109], and found that technical analysis is more profitable in developing markets compared to mature markets. However, their goal is different from ours which seeks to understand if the collective actions of large investor populations can themselves influence the price dynamics of stock markets. Besides, neither their implemented technical indicators, *i.e.*, usually only moving average and its variants, nor datasets are representative.

Self-fulfilling Prophecy. A self-fulfilling prophecy exists where belief in a system or process causes the system to hold true. Whether technical analysis is truly a self-fulfilling prophecy is still under debate; however, it is commonly believed that there exist self-fulfilling aspects of technical analysis [100, 101], *i.e.*, when many investors adopt the same tools and trading rules, they push the price of a stock to move towards the prediction of the signal.

Characterizing the Chinese Stock Market. Researchers are interested in the Chinese market as a representative of stock market in developing countries. [110] states that most Chinese investors are individual investors, who are easily driven by emotion and makes irrational investment decisions. [91] find Chinese markets are still dominated by low-quality information, despite the government's efforts of making better reporting system. These conclusions are consistent with our findings that technical analysis is more prevalent in the Chinese market. Aslo, [111] observes a "more efficient" Chinese stock market following the reform of state-owned-enterprises (SOE) in 2005.

4.8 Discussion and Conclusions

Our work seeks to understand the role of collective behavior in stock markets. We hypothesize that when financial data is not easily accessible, rational investors rely too heavily on common tools that produce synchronized signals for investment decisions. Using US markets as a baseline, our analysis of historical prices in the Chinese markets supports this hypothesis, and shows that it can be exploited to produce returns that significantly outperform even when the market itself are stagnant.

Our work support the view that despite reforms, Chinese stock markets are still inefficient relative to mature markets such as those in the US. These results underscore the need for continued reform, as Chinese market regulators continue to struggle to rein in volatility in the aftermath of the crash of 2015 [112]. In ongoing work, we are applying detailed analysis to other emerging markets, with the goal of more accurately understanding and modeling market inefficiencies over time.

Chapter 5

Gender Bias in Job Market

5.1 Introduction

Despite recent strides made by women in the workplace, workplace inequality persists [10, 11]. In addition to the widely distributed reports of wage inequality across genders [12], it is also well known that there are significantly fewer women in male dominated positions, across both industry sectors (*e.g.,* technology [13]), and job types (*e.g.,* corporate executives [14]).

A number of contemporary theories have hypothesized the source of these gender imbalances, whether they come from different educational paths and biases introduced early on [113, 114, 115], or whether they are the result of workplace attrition [116, 117, 118]. The job search and hiring process might be another contributor [119, 120, 121, 3, 122, 17, 123]. Hidden biases often creep into job listings [124, 125, 18, 126] , and can either actively or passively discourage certain applicants from applying.

The goal of this work is to understand the role that job postings play in introducing or exacerbating gender imbalance in the workplace. More specifically, we are interested in two general questions. First, how significant is gender bias in today's job listings? How

much does gender bias vary over different industry sectors, and how has it changed over time? We hope to track levels of gender bias in job postings through time, to see if any changes are reflected as a result of society's growing awareness of gender bias. Second, we look to measure the end-to-end impact of gender bias in job posts on the actual decisions of potential job applicants. Are applicants aware of such bias in job posts, and do these biases play a role in their decisions to apply for jobs?

To answer these questions, we perform a study with two components, an empirical, data-driven component that quantifies the presence and magnitude of gender bias in job postings over the last 10 years, and a qualitative user-study that seeks to understand the end-to-end impact of biases on whether applicants apply to a posted position.

On the empirical side, we analyze 17 million job posts collected over the last 10 years (2005–2016). We obtain this dataset of job posts through a near-complete download of job posts maintained by LinkedIn, the largest professional networking site online with 500 million users. To quantify gender bias over large datasets, we develop two scalable algorithms that match the same metrics as online services that evaluate job postings for gender bias, Textio and Unitive. We tune our algorithms and show they can approximate Unitive and Textio in bias classification, generating a raw score and a normalized score of gender bias between feminine and masculine. We use a test sample of key words/phrases to validate our approaches against Unitive and Textio. We then apply our algorithm to the LinkedIn dataset to quantify gender bias in the whole market, specific sectors, and its changes over time.

In our user study, we augment our data analysis with a user survey that captures how gender bias in wording actually affects job applicants. We ask detailed questions to 2 user populations, 469 Amazon Turk workers, and 273 undergraduate college students, to understand their perceived levels of gender bias in our job posts, its correlation to specific gendered keywords or phrases, and the ultimate effect they have on the job application

decision.

Our analysis generated several key findings.

1. There is significant gender bias in job listings, but bias has been dropping significantly over the last decade, led by specific job sectors which now trend feminine.
2. Changes in bias levels vary significantly over different sectors, driven by significant changes in usage of a small number of heavily gendered keywords.
3. Our user study shows that users do indeed detect gender bias in job postings, consistent with bias detected by our algorithms.
4. Observed bias still had low levels of impact on user decisions to apply or not apply for a specific position, and there was more sensitivity to bias by men than women.
5. Surprisingly, we observed that users had strong internal biases which played significant roles in their decision on whether they would apply. These biases played a much bigger role than any gender bias language we observed.

To the best of our knowledge, our study is the first large-scale study to look at longitudinal shifts in gender bias in job postings. While our study has clear limitations (lack of historical advertisements from non-LinkedIn job sites, and potential sampling bias in our user study), we believe our results shed light on an important component of the debate on gender equality in the workplace.

5.2 Background

Understanding the evolution of characteristic gender differences in job listings helps to mitigate the effects of masculine and feminine stereotypes, thus reducing gender bias. Here, we provide background on gender disparity in the workplace, and previous efforts to detect gender bias in language.

5.2.1 Gender Equality in Job Market

Historically, certain industries have been dominated by one gender over the other. Approximately one-third of men and women work in occupations with a workforce comprised of at least 80% males or females, respectively. In the past, men tended to dominate engineering and construction occupations while women consistently dominated clerical assistant and teaching occupations [127, 128]. Although census data shows an increase in overall participation of women in the workforce throughout recent decades, the disparity among genders across particular industries remains over time [129].

One reason for such disparity lies in people's stereotypes of genders and occupations. Research shows that people are most likely to seek out occupations that are compatible with one's sense of self [15], suggesting that people are less likely to seek out occupations in industries dominated by the opposite gender. Moreover, a study of perceptions across 80 occupations found that people assume stereotypes associated with the gender of a worker must correlate with the requirements of their occupation. In particular, both genders perceived that masculine physical and agentic qualities were associated with more prestige and earnings [16]. Across industries, managerial positions have historically been perceived as requiring masculine traits. Men even view women negatively for displaying masculine traits in a management role because it is regarded as inconsistent with female role expectations [130]. In the field of Information Technology (IT) and Information Systems (IS), general stereotypes skew towards masculine traits, due to men consistently dominating the field [131, 132].

Another source of gender disparity is institutional discrimination in job markets, which has been observed in both traditional [3, 17, 122, 123] and online settings [126]. In a lab experiment that simulated a hiring decision process, male participants displayed a strong tendency to choose male candidates, even if a female candidate appears as a

slightly better performer [3]. In another field study, comparably matched men and women are sent to apply for jobs in restaurants, and the study found female applicants were significantly less likely to get an offer from high-end restaurants [17]. Later, a similar study was conducted in a male-dominated occupation (engineer) and a female dominated occupation (secretary). Results show statistical significant discrimination against women in the male-dominated occupation and against men in the female-dominated occupation [123].

Another line of research shows that significant improvement in gender equality has been made over of last two or three decades [133, 134, 12]. The overall wage gap between two genders has declined considerably [12], and no institutional discrimination can now be observed in academia [133, 135].

5.2.2 Detecting Gender Bias by Word Analysis

Gender stereotypes are embedded in language use, *i.e.,* different word usage patterns when writing about males or females. Significant prior research used text analysis to examine gender differences in a number of contexts. Some examined how men and women use language differently in text and conversation [136]. Other work studied how text analysis algorithms express unintentional bias, and detect occupational stereotypes in text, *i.e.,* suggesting sexist analogies such as men are analogous to computer programmers and women analogous to homemakers [137]. Other work showed when writing recommendation letters for faculty positions, more standout adjectives are used to describe male applicants than female applicants [138], and women are described as more communal and less agentic (assertive or competitive) [20]. In the context of job advertisements, researchers have shown that language used not only reflects such stereotypical views, but also reinforces the imbalance [125, 124], and that a conscious effort toward

gender-fair language can help reduce it [139, 140]. Finally, much of the prior work in text classification rely on a supervised model with pre-labeled datasets, and is summarized nicely in a survey by Aggarwal et al [141].

Stereotypes can be captured by *gendered words* – terms describing socially desirable traits and behaviors of male or female genders [142, 143]. Gendered words are usually extracted from self-reported characteristics through questionnaires given to college students to measure their self-concept and valuation of feminine and masculine characteristics. The Personal Attributes Questionnaire (PAQ [144]) and Bem Sex Role Inventory (BSRI [142]) are two of the most representative questionnaires in early studies. The words extracted from BSRI and PAQ more typically associate females with more *communal* attributes (*i.e.,* gentle, warm) and men with more *agentic* attributes (*i.e.,* aggressive, competitive). Others generalized gendered words into *expressive* and *instrumental* traits [145]. Tying these together, aggregated lists of masculine and feminine characteristics have been compiled from previous studies, particularly through gendered wording in job advertisements [18]. Finally, Donnelly *et al.* found that women in recent years are less likely to endorse traditionally feminine traits in BSRI [24], indicating that gender norms may require an update of the masculine and feminine stereotyped characteristics.

Based on the lists encoded by Gaucher *et al.* [18], online services like Unitive[1] and Textio[2] use the words and phrases as a baseline to develop gender-neutralizing algorithms with help from machine learning classifiers. The algorithms are trained on internal application and hiring data, and aim at finding gendered wording in job advertisements before recruiters post online. These services represent the state-of-the-art for identifying gendered wording in job advertisements. Since these services run commercial, proprietary algorithms, it is cost prohibitive to evaluate our large job post dataset through

[1] `http://www.unitive.works/`
[2] `https://textio.com/`

their services. Instead, we designed our own algorithms using similar methodologies, and validate them against these online services using samples of test data.

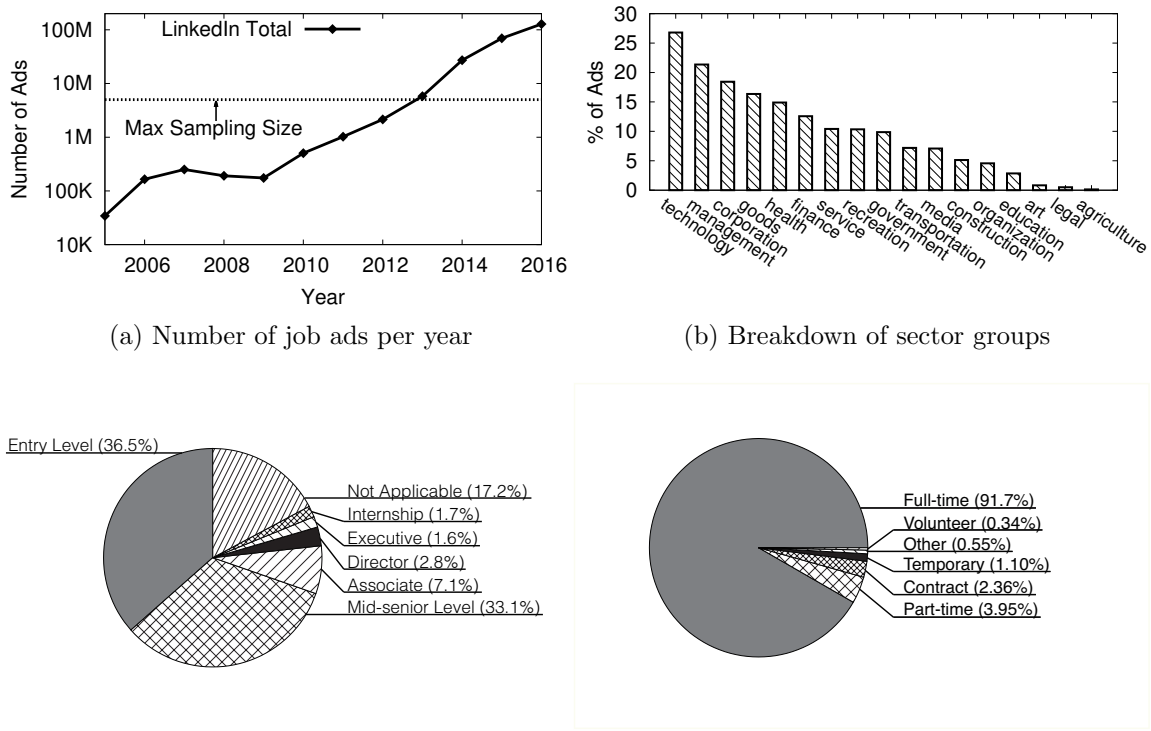
5.2.3 Comparisons to Prior Work

The focus of our study is using large-scale data analysis to characterize gender bias across a comprehensive, longitudinal dataset. Our work was initially motivated by [18], which studied 4,000 job listings (most in a university setting) to characterize gender bias in job listings as an institutional-level mechanism of inequality maintenance. In contrast, we broadly characterize gender bias at scale, using a large dataset that consists of 17 million online job ads covering more than 140 industries. Our work also focuses on examining shifts in gender bias over ten years, and the impact it has on potential applicants' decision.

More recent work [126] identified gender/race discrimination on (performance) reviews in the online freelance (gig) marketplace, by correlating the review with the worker's gender and race. While their work targets reactions to worker output, ours focuses on job advertisements written by only the employer. While [126] analyzed keywords in review, their analysis was limited to sentiment analysis that identifies the attitude of the review, not gender bias.

5.3 Data and Initial Analysis

We describe our data collection methodology and datasets, conduct preliminary analysis on our data, and present basic statistics to provide context for further analysis.



(a) Number of job ads per year

(b) Breakdown of sector groups

(c) Breakdown of employment type

(d) Breakdown of seniority level

Figure 5.1: Statistics of the LinkedIn dataset, 2005–2016.

5.3.1 Data Collection

We collected a large sample of job advertisements from LinkedIn job posts over 10 years (from 2005 to end of 2016). LinkedIn job advertisements are fully public, and accessible online to any user without requiring account registration with LinkedIn. To retrieve a job advertisement, we simply queried known URLs and downloaded the webpages. LinkedIn keeps job advertisements available online for browsing even after the application window has closed. This allowed us to collect a significant longitudinal job advertisement dataset.

Job advertisements on LinkedIn are each assigned a unique ID, which increases monotonically over time. By the end of 2016, the maximum possible ID on LinkedIn reached above 253M, which means there are at most 253M job posts on LinkedIn. Since we

limited our online query rate to avoid overloading LinkedIn's online services, we did not crawl all 253M job postings. Instead, we randomly sample 5 million job post IDs from each year, and only fetch job advertisements matching these IDs. Note that job listings in each year before 2013 contained less than 5 million ads. For these years, we fetched all available job advertisements. After filtering out job advertisements in languages other than English, our dataset contains 17,376,448 job advertisements in total.

Each job advertisement contains a job title, company name, company location, and the main descriptive content of the advertised position. In addition, LinkedIn also provides metadata, including job industry, job function, employment type, and seniority level. LinkedIn has 147 unique job industries[3], which are then further mapped into 17 sector groups, and 35 job functions which describe what activities a person is undertaking. Employment types includes 6 categories: full-time, part-time, temporary, contract, volunteer and other. Seniority level indicates the rank of the position within the business, ranging from entry-level (lowest) to executive (highest).

5.3.2 Preliminary Analysis

Number of Job Advertisements. We plot the number of LinkedIn job advertisements posted per year in Figure 5.2.3(a), as inferred by the total number of possible job IDs in LinkedIn matching a given year. For years up to 2013, our dataset closely follows that of the plotted LinkedIn results, with a small number of missing posts due to non-English listings and unavailable data errors for some of the oldest job postings (likely due to corrupted data). For years 2013–2016, we limited our sample set to 5 million postings per year. While our dataset captures only a limited sample from 2013 onwards, we believe a randomized sample set of 5 million ads is sufficient to capture a

[3] `https://developer.linkedin.com/docs/reference/industry-codes`

representative sample of job postings in any given year.

Sector Groups, Employment Type and Seniority Level. Next, we plot distribution of important metadata fields in Figure 5.2.3(b-d). Figure 5.2.3(b) shows the number of job posts in different job sector groups[4]. We found significant variation among the sizes of different groups: over 25% job posts belong to the largest group (technology) while less than 1% job posts in smallest group (agriculture). As for employment type (Figure 5.2.3(c)), most (91.7%) job listings seek full-time employment, while the rest are mostly split by part-time, contract and temporary (*i.e.,* seasonal). After 2013, LinkedIn introduced Volunteer as a new job sector, which accounts for a negligible portion of our total dataset. For seniority level (Figure 5.2.3(d)), we found a trend of fewer number of applicable job advertisements at higher levels of seniority. This matches our intuition about hierarchies in the job market.

5.4 Quantifying Gender Bias

Our goal is to perform a large-scale analysis of the presence of gender bias over a large corpus of job listings covering the last 10+ years. Our first task is to develop a scalable algorithm to accurately quantify gender bias. In this section, we describe how we emulate the gender bias detection algorithms of two state-of-the-art recruitment assistance services, *Textio* and *Unitive*. We validate our approach by comparing our results against theirs on a small sample of 8,000 job ads. We follow up these results in the next section with a confirmatory user study.

[4]If a job belongs to multiple groups, we count the job in each group it belongs to.

5.4.1 Gender Bias Detection Algorithms

To develop a scalable gender bias detection algorithm, we start by developing metrics to accurately capture different aspects of gender bias. For guidance, we look to the two state-of-the-art recruitment assistance services that measure gender bias, Textio and Unitive. Textio and Unitive are the two largest web services today designed to help potential employers write better job advertisements. Each company curates their own algorithm to calculate whether a given job advertisement expresses masculine, feminine, or gender neutral language. The algorithms draw from an established baseline starting with gendered word lists [18]. We observe and adopt the metrics used by these two services, which we refer to as *Gender Target* and *Gender Tone*. Gender target follows Textio's methodology, which measures the intended audience gender reflected by a job advertisement, and falls into the range of -1 to 1, where -1 means the advertisement specifically targets male applicants, 1 means the advertisement specifically targets female applicants, and 0 means no gender preference is detected. Gender tone follows Unitive's methodology, which captures the extent to which a job advertisement is feminine- or masculine-phrased. It captures a cumulative effect, thus has no fixed range. A gender neutral advertisement has a tone of 0; the more masculine traits stated, the more negative the tone score is, and the reverse for feminine traits. We use the term *gender score* to refer to both metrics.

Gender target highlights feminine and masculine language in job listings. We calculate gender target by first calculating the number of gendered words, with feminine and masculine words canceling each other out. Job ads containing more feminine words than masculine words are considered to be targeting a female audience, and a final score is calculated by applying a sigmoid function on the remaining word count. The same procedure applies when masculine words outnumber feminine words, except that the

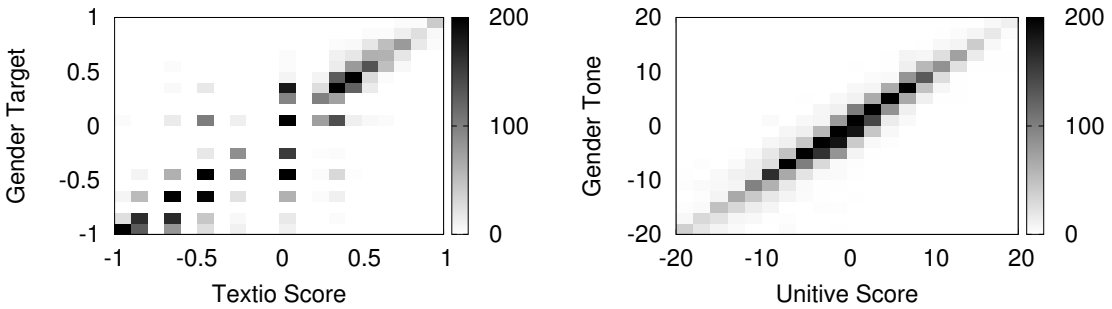
result of this sigmoid function is reversed to fit into the range of -1 to 0. Finally, when the job ad has a perfectly balanced word count, it is considered to be gender neutral, with gender target score of 0.

In contrast, when calculating gender tone, we first categorize terms as *inclusive* (appealing) or *exclusive* (problematic). Prior research [18, 130] has shown a direct correlation with feminine bias from inclusive language, and between masculine bias and exclusive language. In addition, instead of simply counting words, we weights each words based on how gender specific they are. For example, the word "guy" carries a stronger gender implication than "ambitious." Thus, before calculating a cumulative score, gendered words are assigned weights depending on strength of their tone. A strongly masculine word has a strong negative weight, whereas a slightly feminine word has a weakly positive weight. We then add up the weights of all gendered words used in the ad.

Our first key challenge is obtaining an up-to-date list of biased words. To begin, we extracted 50,000 words with the highest frequency from all English LinkedIn job posts we collected, which cover 97.2% of all word occurrences. Since both services highlight words we can categorize with feminine or masculine bias, we queried the services with these words embedded in text, and examined the feedback. Textio annotated 296 gender-related key words, 150 masculine and 146 feminine. We also obtained 843 weighted keywords along with their weights annotated by Unitive, 445 with positive weights (feminine tone), 398 with negative weights (masculine tone). Note that since these two services picked their keyword independently, only 102 words overlap across services.

5.4.2 Algorithm Validation

We validate how well our techniques emulate these services, by comparing our results against theirs. We randomly selected 8,000 job advertisements from our dataset, and



(a) Textio score and Gender Target (b) Unitive score and Gender Tone

Figure 5.2: Comparison of actual and approximated gender-bias metric.

uploaded them to Textio and Unitive using their free online accounts. We compared the results they return to results from our own algorithms. We plot our results against results given by Textio and Unitive in Figure 5.2. For Textio, the scores are more scattered due to the use of discrete count before normalization. We found that 71.8% of the gender target scores are within a difference of 0.1 from Textio scores (an error rate of 10%). In the case of Unitive, the scores match along a straight line with a slight bias of -0.209 towards masculine tone, and 77.5% of the scores have an error of 1 or less. Since Unitive scores varied by as much as 10, this also represents an error rate of roughly 10%.

The error in our scores is due our inability to recover full keyword lists from both services, especially for phrases. Given the highly subjective nature of gender bias, our goal is not to generate "perfect" algorithms, but to obtain general and scalable algorithms with results that approximate public systems.

5.5 Longitudinal Analysis

In this section, we apply our gender bias detection algorithms on to our LinkedIn job post dataset. Our results show that in recent years, wording in job advertisements skews masculine, but the absolute level of bias is becoming more neutral. First, we identify a

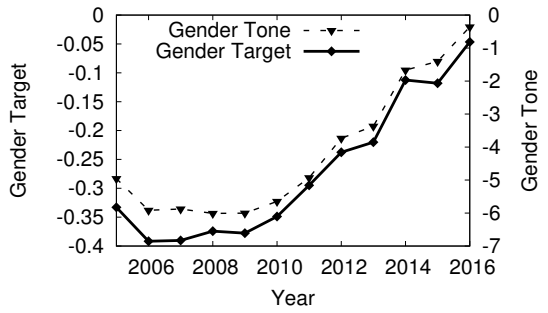


Figure 5.3: Average gender bias score from 2005 to 2016.

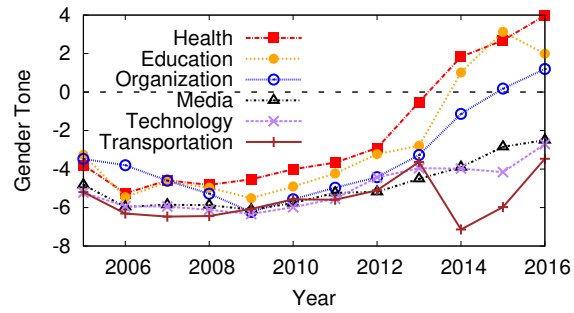


Figure 5.4: Gender tone over time of top 3 and bottom 3 sectors in 2016.

few factors that contribute to the trend. Different job functions across industry sectors distribute unevenly in terms of gender score. Although this uneven distribution of jobs across sectors results in an overall averaging of feminine and masculine bias scores, the effect is limited. Second, the masculine bias comes primarily from formal and long-term employment jobs, and appears more severe in senior level positions. Over the 11-year period, the number of entry-level jobs posted is increasing over time, which partially accounts for the decreasing masculine bias, as these positions predominantly skew feminine. Third, to quantify the effect of these factors, we formulate a regression analysis to predict gender score, which shows that the effect of all factors are significant. However, there is still an underlying trend of decrease masculinity after separating out the effect of these factors, indicating possible increasing awareness of using more gender neutral language. Finally, we identify a few gendered words that contribute the most in driving change in levels of gender bias.

5.5.1 Gender Bias Over Time

We begin by studying how gender bias in job postings changes over time. We are interested in whether the market as a whole (and perhaps as a proxy for the general population), is becoming more aware of gender bias. We compute two values for each

job posting: a *gender target* score and a *gender tone* score. For each year, we compute the average scores and standard deviation of all job postings collected from that year. Figure 5.3 uses a dual Y-axis to compare the average score of the two algorithms. The standard deviation values are similar over time and the two algorithms, thus they are omitted.

We make some key observations. First, the average gender scores remain consistently below 0 across all years, indicating that the job market, as captured by LinkedIn postings, is skewed towards masculine appealing positions. Second, an increasing absolute score over time suggests that the market is becoming more gender neutral. Third, while our two metrics use very different algorithms and their absolute scores are not directly comparable, their trends over time are almost identical. We performed another consistency check of these results using the gendered word lists from prior work [18], and the results are highly consistent. The frequencies of the three trends show strong correlation between each other (p-value < 0.0001), with Pearson correlation of more than 0.97. This confirms that the overall trend is fundamental to the job market, and the two metrics capture consistent views of the same phenomenon over time. To get a better understanding of where the masculine jobs originate and to explain the trend over time, we explore a variety of dimensions to better understand the underlying structure of the LinkedIn job market.

5.5.2 Gender Score over Job Sector Groups

We dive down to see how individual sector groups are changing over time with respect to gender bias. Recall that all together we have 147 distinct industries, mapped to 17 sector groups. While we have established that bias is decreasing over time for the entire job market, we want to observe any variance in dynamics across different job sectors.

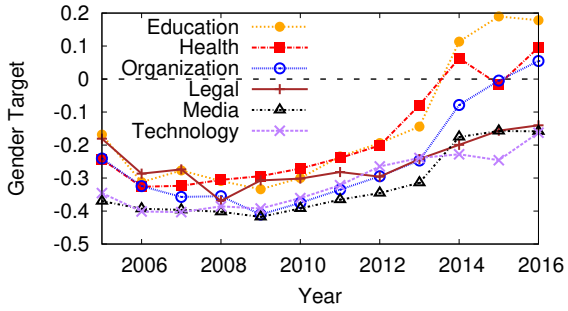


Figure 5.5: Gender target over time of top 3 and bottom 3 sectors in 2016.

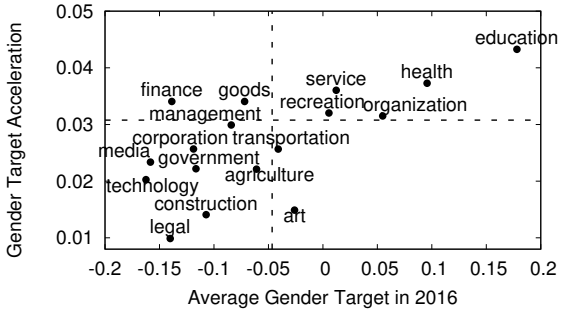


Figure 5.6: Gender target in 2016 vs. gender target acceleration across sector groups.

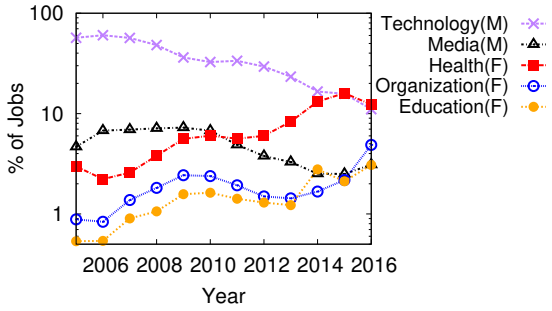


Figure 5.7: Shift in distribution of 2 typically masculine and 3 typically feminine sector groups.

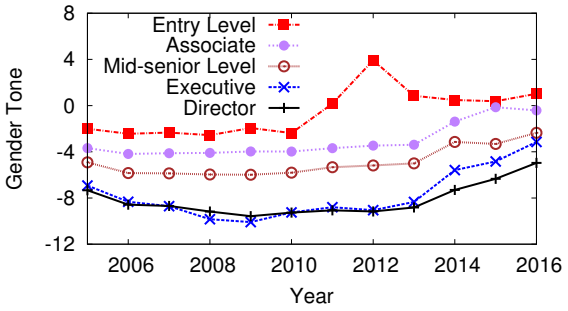


Figure 5.8: Average gender bias score of different seniority level (*gender tone*).

We begin by looking at the top and bottom sectors sorted by gender scores. In Figure 5.4 and Figure 5.5, we plot the top 3 and bottom 3 sectors sorted by 2016 gender tone and 2016 gender target scores respectively. For each sector, we trace back their scores over past years. First, we note that gender tone and target are remarkably consistent in their top and bottom sectors. In both cases, Education, Health, and Organization are top sectors (more feminine), and all have risen consistently over time to current values above 0 (the dashed line represents the value 0). Media and Technology are the two sectors that appear as bottom sectors in both metrics. Their scores are rising, albeit at much slower rates, and occasionally experience short term dips (the start of the great recession 2008–2009). The Tech sector also experiences another dip around 2013–2015, showing

that perhaps the most biased sectors might be more sensitive to economic downturns.

In Figure 5.6, we plot each sector's *acceleration* of gender scores over time, against its 2016 gender target value. Acceleration is computed as the slope of a linear regression of a sector's scores over time. The results are intuitive. The sectors slowest to reduce masculine bias (Tech, Legal, Construction) still have some of the most masculine biased gender target scores in 2016. Others like Education and Health have high rates of change towards more feminine wording, and as of 2016, are firmly on the side of feminine bias.

Dynamics of sector groups are correlated with the overall increasing gender score. One reasonable question is how much dynamics between sectors contribute to the overall gender bias trend. To answer this, we first calculate the ratio of gendered job postings inside each job sector. We find that the number of jobs is increasing in several stereotypically feminine sectors and decreasing in a number of stereotypically masculine sectors (see Figure 5.7). So, it is possible that the overall increasing gender score comes from a changing of sector distribution. To remove such effect, we recalculate the gender score across the entire 11-year period, but weigh each sector based only on the 2016 distribution of jobs across sectors. The result shows that the impact of shifting weights across sectors is small: gender tone only increases at most 0.56 under the new calculation (and gender target only increases by 0.034), much smaller than the overall increasing trend showed in Figure 5.3.

5.5.3 Gender Score over Seniority Levels

Next, we break down all job postings by their *seniority level*, and compute average scores in each category. Figure 5.8 shows the breakdown of results for different seniority levels, where there is a clear correlation between seniority ranking and the masculine tone of the job posting. We omit results of gender target here, since they show very similar

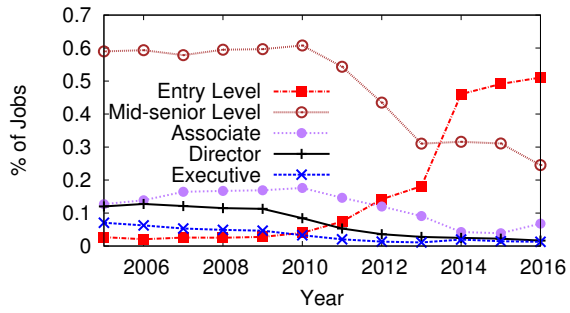


Figure 5.9: Distribution of seniority level from 2005 to 2016.

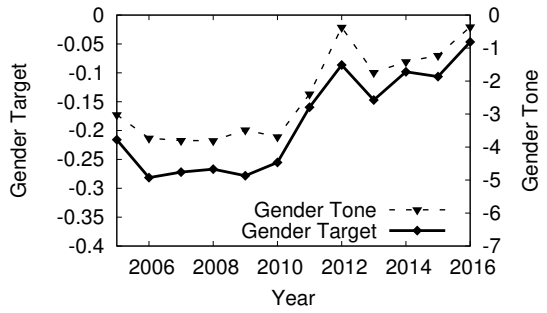


Figure 5.10: Average gender score computed over distribution of seniority level in 2016.

trends. These results are consistent with prior work [130] that discusses how men hold an overwhelming majority of top management positions, and thus masculine traits are commonly associated with these higher-ranking positions. Due to a phenomenon called *ambivalent sexism* [146], attitudes regarding gender roles presume women historically belong in a domestic setting and are incompetent at holding positions of power. These unconscious biases may persist today, and are likely used to explain the gap in gender participation rates at more senior level positions.

Similar to trends across sector groups, we also find the distribution of different seniority levels changes over time (shown in Figure 5.9). It is clear that the increase in number of entry-level jobs corresponds to a decreasing proportion of mid-senior level jobs over time. Since entry-level jobs tend to be less biased towards masculine, the shift in distributions affects gender score. In Figure 5.10, we show the effect of removing this factor by computing the gender score using a fixed seniority level distribution from 2016. Compared to Figure 5.3, we get a similar increasing score trend, with a much smaller magnitude of masculinity. Thus, we conclude that the overall increasing trend comes from two parts: increasing lower-level jobs and increasing feminine language in each seniority level.

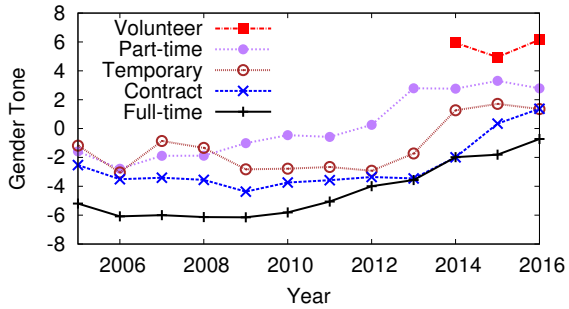


Figure 5.11: Average gender bias score of different employment types (*gender tone*).

5.5.4 Gender Score over Employment Types

Figure 5.11 shows how the gender score is distributed over different employment types with respect to *gender tone*. Gender scores show clear and consistent trends in different employment types, and the more formal and long-term the employment, the more masculine tone in the job posting. Since over 90% of jobs are full-time jobs across all years of our dataset, we do not investigate the effect of changing distribution in terms employment types.

5.5.5 Comparison of Gender Bias Contributors

Finally, after observing how these different factors affect the job market, we aim to quantify the effect of each factor. To do so, we formulate a regression analysis. We use seniority level, year, sector group and employment type as independent variables, to predict the gender score of a job advertisement. The reference categories for seniority level and employment type are "N/A" and "other," respectively. Since a single job can belong to multiple different sectors, there is no redundancy in sector groups.

Our result is shown in Table 5.1 and Table 5.2, for gender tone and gender target, respectively. We find that all the independent variables have statistically significant effects on gender score. The effects are consistent with the previous qualitative analysis,

Feature	Coefficient
(Intercept)	-1.1380
Year after 2005	0.3064
Entry Level	2.9893
Associate	0.6209
Mid-senior Level	-0.6825
Director	-4.6385
Executive	-3.1562
Volunteer	2.4624
Part-time	-1.8491
Temporary	-2.7402
Contract	-4.4472
Full-time	-5.0084
Transportation	-3.6283
Technology	-1.5921
Media	-1.0557

Feature	Coefficient
Agriculture	-1.1434
Construction	-0.9402
Management	-0.3033
Finance	-0.3763
Art	-0.0245*
Organization	0.2408
Corporation	0.2867
Legal	0.3372
Recreation	0.4573
Service	0.8327
Government	1.2678
Goods	1.6498
Education	2.4565
Health	3.9049

Table 5.1: Ordinal regression using *gender tone* as dependent variable. $p < 0.001$ applies for all entries except *, which has $p = 0.273$.

i.e., gender scores vary over groups, and decrease with higher seniority level and more formal employment. However, after ruling out the effect of these factors, we still find an underlying increasing trend that is statistically significant. Although there could be other factors, we believe that awareness of using more inclusive language, and/or using less masculine language, is an important part of the change.

5.5.6 Changes in Word Use

Finally, we are interested in understanding how different words and phrases vary in their contribution to gender bias over time in job listings. We take the 500+ gender biased words from our dictionary, and plot their frequency of appearance (and therefore impact on gender scores) in Figure 5.12. We find that the frequency distribution of the most popularly used terms is exponential (and therefore it appears linear on a log plot).

Figure 5.13 plots the frequency of usage for top 20 words across the years, where

Feature	Coefficient
(Intercept)	-0.0510
Year after 2005	0.0208
Entry Level	0.1000
Associate	-0.0193
Mid-senior Level	-0.0885
Director	-0.2275
Executive	-0.1589
Volunteer	0.0442
Part-time	-0.1497
Temporary	-0.1583
Contract	-0.2670
Full-time	-0.2848
Technology	-0.0731
Media	-0.0620
Agriculture	-0.0571

Feature	Coefficient
Transportation	-0.0458
Corporation	-0.0432
Legal	-0.0414
Construction	-0.0303
Finance	-0.0243
Management	-0.0156
Government	-0.0041
Service	0.0150
Organization	0.0203
Goods	0.0328
Recreation	0.0401
Art	0.0621
Health	0.1345
Education	0.1974

Table 5.2: Ordinal regression using *gender target* as dependent variable. $p < 0.001$ applies for all entries.

we group the masculine and feminine words separately in the figure. We see that most of the masculine words show a stable or slightly decreasing trend[5], while the feminine words display more dramatic changes over time. Specifically, the most used masculine word, "strong", experienced a sharp decrease, potentially due to the growing awareness of biased wording. Among the top feminine words, "care," "patient" and "health" display significant increase, possibly driven by the growth in health-related jobs, while "understand," "develop" and "relationship" show a visible decline over the years. We also repeated our study on words with the largest changes in usage frequency, which produced similar results.

Finally, we compute the frequency of usage for each word across the years, and fit a trend line using linear regression. Of these terms, 145 words show statistically significant trends (p-value < 0.05). 75 of these are masculine toned words, and they are evenly

[5]The only exception is "driver." This is probably because driver can also be a job title in the transportation sector which experienced a rapid growth since 2014.

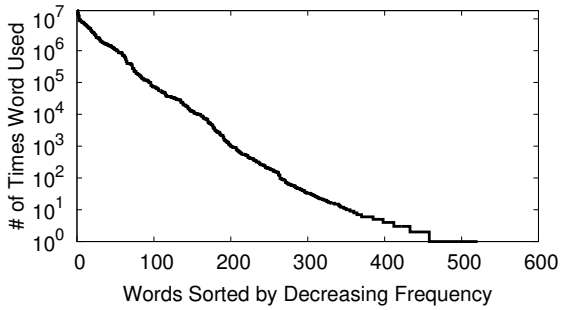


Figure 5.12: Distribution of word frequency.

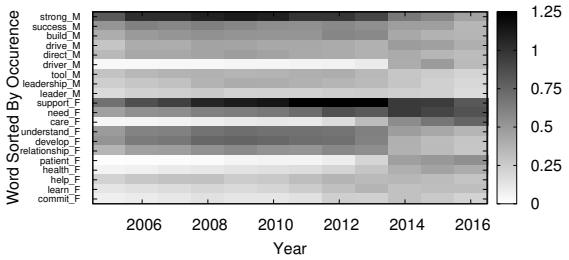


Figure 5.13: Evolution of word frequency for top 20 words.

divided between those growing in frequency and those dropping in frequency. Of the remaining 70 feminine toned words, the large majority (84%) showed an increasing trend.

5.6 User Study: Impact of Gender Bias

So far, we have quantified the level of gender bias in job listings over time, but we do not yet understand how these gender biases actually impact users (potential job applicants). To answer this question, we conducted a user survey, which we describe here. In short, we find that gender scores from our algorithms properly reflect perceived gender stereotypes associated with job postings, but that biased wording has limited effect on the perception of a job, compared to respondents' preconceived notion of the job type. We also find that male respondents are less willing to apply for stereotypically feminine jobs, while the reverse does not hold.

Survey Participants. We recruit survey respondents from two different sources, Amazon Mechanical Turk (MTurk), and undergraduate students from UC Santa Barbara. In total, we received results from 469 distinct MTurk workers and 273 students, each job advertisement is evaluated by at least 20 different workers and 12 different students. Undergraduate students were volunteers who received necessary credit for their course

work. Each worker was compensated $1 for finishing the task. To ensure the quality of replies, we require workers to have an 80% HIT approval rate, and have at least 50 HITs approved in the past. We also include a quality check (*i.e.,* gold standard) question in our survey question list, (*e.g.,* "Please answer A and D for this question.") to avoid low-quality/non-responsive workers. For respondents who failed our gold standard questions, their responses are not included in our analysis. In most cases, responses from MTurk workers and the students point to the same conclusion, and we thus combine their answers in such analysis. In cases when the responses differ, we analyze the respondent pools separately.

The demographics of our survey participants are as follows. Among all 469 Mturk workers, 54.6% indicated male and 45.4% indicated female. The majority of participant ages fall into the ranges of 21 to 30 (38.8%) or 31 to 40 (34.3%), with 1.49% younger than 21, 11.7% older than 50, and the rest fall between 41 and 50. Most participants work full-time (67.2%); and most hold a Bachelor's (42.9%) or a Master's (24.5%) degree. Among the 273 college students, 201 (73.6%) are female and 72 (26.4%) are male. 209 (76.6%) of the students are younger than 21, and 64 (23.4%) are from 21 to 30.

Methodology. In our user study, we divided job advertisements into 3 categories: *masculine jobs*, *feminine jobs*, and *gender neutral jobs*. Feminine jobs are randomly sampled from job advertisements with the highest 10% gender score, as scored by both gender target and tone. Masculine jobs are similarly sampled from postings with the lowest 10% gender score, and neutral jobs are sampled from advertisements with scores nearest 0. We did not restrict the timespan of our advertisements, since we want to maximize user reaction on potential gendered language. When older job advertisements is included in our sample, we manually check their content to make sure there are no outdated words that will significant affect user's reaction.

For each advertisement, we created a second version of the job description by re-

placing keywords or phrases marked as gendered language from our dataset with more gender-neutral words or phrases not in our dictionary of gendered words. For example, substitutions included "workforce" replaced by "employees," and "collaborating" replaced by "working." Although we made efforts to consistently replace biased words with the same neutral alternative, some instances required more dynamic replacements to retain the readability and intent of the original post. For instance, depending on context, "engage" may be replaced with "participate," "employ," or "work." Other words, like "please," were simply removed. To ensure the biased language was not just replaced with different biased language, we calculated gender scores for the edited descriptions, and verified that the substituted descriptions received gender-neutral scores.

We replaced or removed as much gendered words and phrases as possible without changing the intended meaning of the original job posting. This provides a suitable baseline to isolate the impact of gender wording from people's inherent biases and stereotypes. This contrasts with prior studies [18] that analyzed masculine and feminine language without comparing against neutral wording as a baseline.

We asked each user to read three job advertisements, one from each category. After reading through the ads, we gathered their responses to the following questions:

- *Q1*: If you were fully qualified to apply for a job like this, how likely is it that you would apply for this particular position? Answers are measured by 5-level Likert Scale (1 indicates definitely would not apply and 5 indicates definitely would apply).
- *Q2*: By looking at the job description, what would you think to be the percentage of women currently working in this type of position?
- *Q3*: While reading the job description, to what extent did you feel that the advertisement would attract more male or more female applicants? Answers are measured by 5-level Likert Scale (1 indicates job attracts mostly males and 5 indicates job attracts

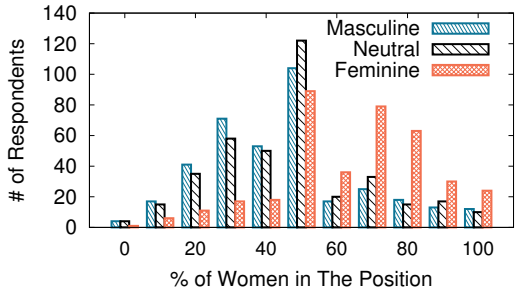


Figure 5.14: Responses on Q2 (percent of female in the position).

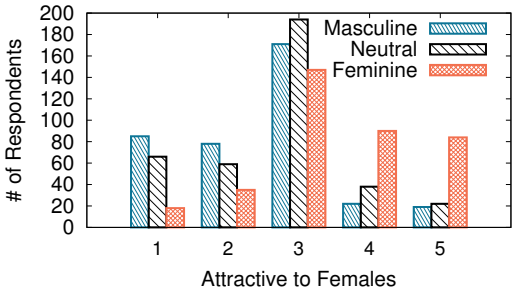


Figure 5.15: Responses on Q3 (attracting female applicants).

mostly females).

- *Q4*: Please mark any words or sentences that you do not feel comfortable with.

Q1-Q3 in the above questions are multiple-choice, and Q4 is open-ended. At the end of the survey, we collected user demographic information. We consulted our local IRB[6] and obtained approval before conducting the user study. Note that while Q3 could be asked differently, *i.e.,* , ask users to rate the attractiveness of the job and compare results between male and female respondents, we chose this version so users would focus on the effect of wording rather than allowing other, random or uncontrolled factors to influence their "broad" evaluation of a post. We attach a full survey script in Appendix A.2.

The three job advertisements were randomly selected from the three categories (masculine, feminine, neutral), with equal likelihood of choosing an edited or raw version for each category. We used a pilot test of Amazon Turk users to determine if question order impacted user response. After controlling for other factors, results showed a decreased likelihood of job application for the second and third ads. Thus we presented the three sampled advertisements to participants in random order.

[6]The IRB protocol number is 29-17-0259.

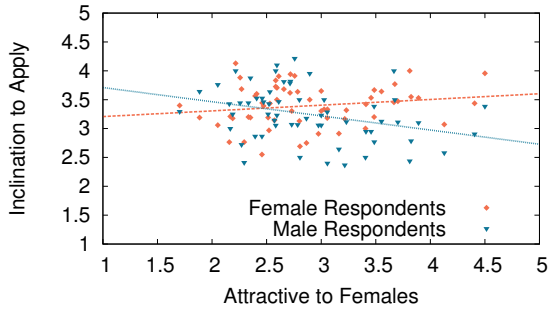
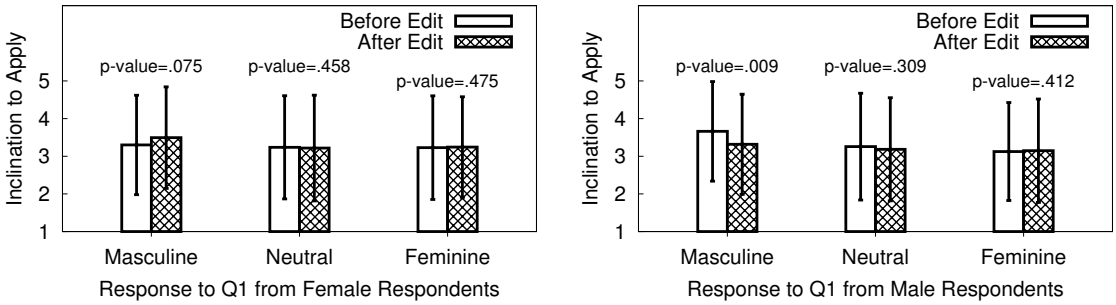


Figure 5.16: Average user response on Q1 against average user response on Q3.

5.6.1 Preliminary Task: Do gender scores accurately reflect user perceptions of gender bias

We first validate our algorithms designs with the user study. Here, we compared user responses to questions Q2 and Q3 for different types of job advertisements before replacing or removing gendered words. The results are shown in Figure 5.14 and Figure 5.15. In each figure, we analyzed responses for each question choice, across all advertised job positions. We find that people presume more female workers in supposed feminine positions, suggesting feminine toned job advertisements appear more attractive to women. Corresponding findings apply to masculine toned job advertisements, as well, thus validating our algorithms.

We also conduct Mann-Whitney U-tests on the distribution of responses between groups, and the difference is statistically significant with p-value less than 0.001 across all types except between masculine and neutral ads. Interestingly, all distributions show an artificial peak at the most neutral answer. Many respondents selected 50% for Q2, where we asked about the percentage of women working in the position, and for Q3, many selected the option "attracting male and female applicants equally." Respondents who selected the neutral choice often provided reasons related to "equal chance" or "equal rights," showing a conscious awareness of, or desire for, equal gender participation in the



(a) Female respondents (b) Male respondents

Figure 5.17: Responses from university students on Q1 (inclination to apply), breakdown by types of jobs. Differences measured with Mann-Whitney U-test.

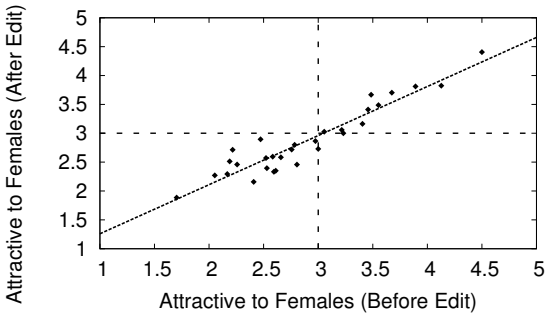


Figure 5.18: Average answer to Q3 (attracting female applicants) before and after word substitution.

workplace.

5.6.2 Principle Task: Quantifying effects of gendered wording on job application rates

Perceived occupational gender bias affects decisions to apply.

Given the correlation between word choice and perceptions of gender bias, we next sought to examine the extent to which this perceived bias influences one's decision to apply. We show that male respondents express noticeably diminished inclination to apply for jobs perceived to predominately attract females. We quantify the level of perceived

gender bias by averaging responses to Q3. In Figure 5.16, we plot the perceived bias of a job against the average tendency of female and male applicants to apply, indicated by average response to Q1. By applying linear regression on both male and female applicants, we discovered that female applicants do not show any preference with respect to gender distribution, with near zero slope (0.0989) and a p-value of 0.190. Meanwhile, male applicants displayed a preference against applying for female-dominated jobs, with a slope of -0.245 and a p-value of 0.0113. This contradicts prior work [18], where female applicants found masculine worded occupations significantly less appealing. One explanation is that our gender bias is naturally embedded in the job posts, and thus likely to be of a lower intensity than artificial job ads composed specifically to contain gender bias. Additionally, female perception of and reaction to gender bias may have shifted since the 2011 study.

From Figure 5.16, we can observe a high degree of variance, indicating that willingness to apply for a job may be affected by other external factors besides gender neutrality. When we asked their reason for why they will or will not apply for a position, we found a few frequently mentioned reasons, including the anticipated salary, benefits, location, workload, potential of career development, and whether the field of job appeals to the respondent.

Changes in gendered wording have limited effect on predisposition to apply.

The ultimate question remains as to whether a recruiter can change the wording in a job advertisement and increase the likelihood of potential applicants to apply for the job. Thus, we seek to quantify the causal effect of wording on users' decisions to apply for a job or not. For female and male applicants, we compared the predisposition to apply for a job, measured by averaging answers to Q1, before and after word substitution.

These results are different between the two pools of respondents. For students, word-

ing change in masculine-worded ads does affect application decisions, as shown in Figure 5.17. Removing male-biased words from job advertisements leads to less male applicants and marginally more female applicants expressing an inclination to apply. When performing Mann-Whitney U-test on responses of MTurk workers, the p-value are above 0.05 for all three job types with both male and female respondents.

This shows that the effects of word use are observable, but somewhat limited. We then sought to break down the effect, pinning down whether wording actually *causes* a perceived bias, by comparing respondents' reported perception before and after word substitution.

We plotted the average responses to Q3 before and after word substitution. If wording is the sole cause for gender bias, then by removing the biased words, all jobs advertisements should appear with similar level of perceived gender bias, thus yielding a slope of 0. In contrast, if wording has zero impact on gender perception, it will show a slope value of 1. In Figure 5.18, we can see that the perceived bias persists even after word substitution, with a linear regression yielding a slope of 0.850 and p-value of 0. Similar results are observed for Q2, showing a slope of 0.825 and p-value of 0. This indicates that there certainly exist other properties affecting gender perception more influential than changes in wording.

Preconceived notions of occupations predominately affect user perceptions.

To better understand what factors influence user perception, we examined the reasons given in our survey responses. In the survey, we asked users to explain their reasoning and mark any words or phrases in the job advertisements that made them feel uncomfortable. With this exercise, we hoped to gain insights into current perceptions that may be missing from previous studies or even current available services.

We found that many explanations given in our responses include preconceived ideas of

the described job function. For example, in response to a job providing technical support for customers in a cable television company, one user believes that 20% of the workers in this position consist of women and therefore presumes the position will attract primarily male applicants, indicating as the reason "It's a technology job." Some respondents even expressed strong gender stereotypes, making statements like, "Low wage jobs tend to hire women, men try to get better jobs." Similar stereotypes also affected users reading posts for jobs perceived to be suitable for female applicants.

Many responses associated a particular gender with specifics characteristics they assumed would best fit the job. For instance, some highlighted phrases such as "bringing accountability, decency, and humor to the job," with explanations stating that these expectations would appeal primarily to male applicants, and women may not like the position. We infer that these users think women are less likely to possess such attributes, making them unfit for the job. These words were not included in our gendered language dataset, indicating modern perceptions of occupational gender bias.

Other responses focused more on preconceived notions of job functions. One response to a position requiring business travel with the company CEO described how they couldn't imagine a man doing this kind of assistant job, demonstrating an inherent stigma against men performing clerical work. Other respondents indicated that non-assistant positions requiring travel or "with little supervision" were better fits for male applicants who may feel more comfortable traveling than women, perhaps due to traditional views that women should or would want to stay at home. Additionally, users suggested that job descriptions requiring an ability to lift up to 50 lbs. or unloading trucks skewed towards male applicants who would be more likely to be capable of such physical activity. On the other hand, many male respondents expressed no interest or consideration for a beauty consultant position because they perceived it as a field of work for females, with some users describing a beauty school degree requirement as simply, "sounds sexist." Most

surprisingly, several responses ironically stated that using the phrase "Equal Opportunity Employer" felt insincere and directly singles out females or minorities.

We originally intended to use these questions to better identify specific gendered words or phrases. Surprisingly, we instead gained insights about the role that inherent gender bias plays in the job marketplace.

5.7 Discussion

Through our data analysis, we observed an increasing shift away from masculine-biased job postings over the years, and that employers today use less gendered wording than they did 10 years ago. However, the results of our user study also indicate that this trend towards gender neutral wording does not correlate with a perception of gender neutrality in the job market.

Surprisingly, user responses to our survey showed significant gender bias in the responders to specific job positions. Despite the correlation we found between gendered wording and perceived bias, users' explanations show their underlying biases were bigger determinants of their likelihood to apply than any gendered wording. Even after removing all gendered language from the job advertisements, these trends remained in the responses (see Figure 5.18). Gender bias is present in the responder's own perception, independent of the language used in job posts. The implication is that completely removing gendered wording will have limited impact in forming a more gender neutral workforce. This echoes observations made in prior work [126] that inherent user bias was pervasive in the job marketplace. Ultimately, we need to address inherent gender bias in the applicants themselves to significantly improve gender neutrality.

5.7.1 Limitations

Our methodology and data source introduced a number of limitations in our study. First, our data analysis shows a trend of increasing gender neutrality over the years, and we examined the impacts from different factors. However, all the observations are based on co-occurrences, so we cannot make claims of strong casual relationships in our high-level results. Second, our job postings dataset is from a single source, LinkedIn. While it is arguably the largest job listing site online, it is still prone to hidden biases, *e.g.,* towards US based positions or more technology sectors. Unfortunately, other large job listing web services like Monster or Indeed do not provide historical job postings for analysis.

For our user study, there are potential biases in our sampling. In our user study we recruited 25 users to evaluate each job advertisement, but we only studied 30 job advertisements in total. A small number of job advertisements may not represent the largely diverse pool of all the job advertisements. In our user study, we are limited to Amazon Mechanical Turk workers and undergraduate college students, neither of which are representative of a highly diverse workforce in the general population. In addition, since the workers do not necessarily evaluate jobs from their own area, some respondents expressed unfamiliarity with terminology (acronyms, corporate jargon) specific to the field of work described. Finally, answers to our questions showing gender bias may in fact be reflecting personal familiarity of respondents with assumed statistics in a given industry.

While quantifying and understanding these limitations will require further studies, we believe our analyses provides an early empirical perspective on the shifting dynamics of gender bias in the American workplace. We hope our work will encourage further studies of large-scale gender bias, and help identify the key factors that will lead to improved

gender quality in the workplace.

Chapter 6

Detect Gender Stereotype in Natural Language

6.1 Introduction

Gender stereotypes are frequently observed in language. Recent studies have identified description that reflect gender stereotype in different types of articles, such as biographical pages of notable people [19], recommendation letters [20], fictions [21] and movie dialogs [22]. The prevalence of gender stereotypes in languages may reinforce the concept of gender stereotypes like aggressive men and domestic women.

Although gender stereotype are frequently observed, few studies try to systematically detect them. Gender stereotypes are traditionally captured by gender word inventory: pre-compiled word lexicon which contains items describing social traits and behaviors that differentiate male or female genders [142]. These words are extended in later studies for detecting gender bias in job postings [18, 25]. Though widely applied, items in traditional gender word inventories are less endorsed by women in recent years [23, 24], and the performance of using these words to detect gender stereotype in language is

unclear.

Meanwhile, with the wide availability of large-scale text data, people start to tackle natural language processing tasks using an end-to-end approach. Instead of using a pre-compiled word lexicon, the end-to-end approach trains a neural network model that produce the desired output using labeled raw text as input. The end-to-end approach has shown great success in tasks such as sentiment analysis [147] or hate speech detection [26], which were traditionally solved by lexicon approach [27, 28].

The goal of this work is to systematically study the problem of detecting gender stereotype in natural language. More specifically, we are interested in two general questions. *First*, can we update the gender stereotype lexicon that reflect gender stereotypes in modern society? Can we use the new lexicon to detect gender stereotypes in natural language? *Second*, can we build a gender stereotype detection model using the end-to-end approach? If so, how is the performance compare to the traditional lexicon approach? How costly it is?

To answer these questions, we perform a study with two components: building a gender stereotype lexicon, and comparing lexicon approach and end-to-end approach.

For building the gender stereotype lexicon, we analyze English Wikipedia data and extract verbs and adjectives that are used to describe human. We select most frequently used words and ask users to evaluate the masculinity and femininity of each word. We then extend the human labeled words to all the verbs and adjectives, result in a gender stereotype lexicon that contains stereotype score of over 10,000 words. For comparison between the end-to-end approach and lexicon approach, we collect a dataset that contains articles that are consistent with or contradict common gender stereotypes. Using this dataset, we develop stereotype detection algorithms using both our lexicon and end-to-end approach.

Our work has the following contributions:

- We develop the gender stereotype lexicon that reflects gender stereotypes of modern society.
- We collect the first human labeled text corpus for gender stereotype detection.
- We formulate gender stereotype detection as classification tasks, and develop a lexicon approach and an end-to-end approach to deal with the tasks.
- We demonstrate that the end-to-end approach significantly outperforms the lexicon approach. The end-to-end approach does not need a large training data to outperform the lexicon approach.
- The end-to-end model we developed can be applied to gender bias detection in job advertisements and it outperforms the industry state-of-the-art.

To the best of our knowledge, our study is the first study tackles the problem of gender stereotype detection. We believe our results shed light on an important component on better understanding of gender stereotypes in language.

6.2 Related Work

6.2.1 Gender Stereotypes in Language

Gender stereotypes are common beliefs about what men and women are like and should be like. According to gender stereotypes, women should display *communal* traits (*e.g.,* nice, caring, warm) and men should display *agentic* traits (*e.g.,* assertive, competent, effective) [45, 46].

Gender stereotypes emerge in language use when individuals freely describe men and women [148]. It has been found that the category label that used to refer to a group automatically activates the characteristics stereotypically associated with the group, even in unprejudiced people who do not endorse the stereotype [149, 148]. This also applies

when the category label is one's gender. For example, after primed by words that are consistent with gender stereotypes (*e.g.,* "nurse"), people are faster in associating gender pronoun (*e.g.,* "she") with correponding gender (*e.g.,* "female") [55].

As a result, gender stereotypes are common in contemporary languages, both in written language and spoken language. For example, in friction writing, traditional gender stereotypes such as dominant men and submissive women are common throughout nearly every genre, no matter whether the author is a man [21]. On Wikipedia, articles about notable women emphasized more on romantic relationships or family-related issues compare to articles about notable men [19]. When writing recommendation letters for faculty positions, women are described as more communal and less agentic than men [20]. In movie dialogs, male characters use more words related to achievement than female characters [22].

The gender stereotypes are frequenly observed in previos works. There is still a lack of systematica study for language stereotype detection. Our work seeks to investigate algorithms that detect gender stereotypes in natural langauge.

6.2.2 Gender Word Inventory

Stereotypes can be captured by *gender word inventories* – pre-compiled lists of items describing social traits and behaviors that differentiate male or female genders [142, 143]. Gender word inventories are usually extracted from self-reported characteristics through questionnaires given to college students to measure their self-concept and valuation of feminine and masculine characteristics. The Personal Attributes Questionnaire (PAQ [150]) and Bem Sex Role Inventory (BSRI [142]) are two of the most representative questionnaires in early studies. The items extracted from BSRI and PAQ typically associate females with more communal attributes (*i.e.,* gentle, warm) and men with

more agentic attributes (*i.e.,* aggressive, competitive), which are highly consistent with common gender stereotypes. Others generalized gendered words into *expressive* and *instrumental* traits [145]. Tying these together, aggregated lists of masculine and feminine characteristics have been compiled from previous studies, particularly through gendered wording in job advertisements [18].

The gendered word inventories are usually shown to people to measure their gender identity, *i.e.,* whether people see themselves as masculine or feminine. Among them, BSRI is considered as a golden standard in gender role detection, and it has been used in thousands of studies in more than 40 years after it was developed [151]. However, it was found that items captured by BSRI are less endorsed by women in recent years [23, 24]. These works reviewed a large collection of studies that apply BSRI, and track how does user responses change over a long period of time. The women's femininity score is decreasing significantly over the years, indicating that gender norms may require an update of the masculine and feminine stereotyped characteristics.

Given previous result showing that the existing gender word inventory may not capture these concepts in the modern world, we seek to develop gender stereotype lexicon that captures people's perception of gender stereotypes in the contemporary society.

6.2.3 Gendered Stereotype Studies in Natural Language Processing

To best of our knowledge, there is no existing tools or algorithms that try to determine if a piece of text is consistent with gender stereotype. One class of tools that resemble ours to some extent is detecors of gender biased language in job advertisements [25]. These tools leverage pre-compiled gender bias words that may affect decisions of job applicants, and calculate gender bias of a job advertisement based on the number of occurrence of

these words.

Although detecting gender stereotype in natural language is still an under-explored area, natural language processing community starts to pay more and more attention to fairness and bias. Most of the existing works focus on identifying biases in algorithms and debias these algorithms. For example, word embeddings make stereotypical analogies such as "man : computer programmer :: woman : homemaker" [137]. Previous studies also observed performance discrepancy across genders in systems including coreference resolution [152], image caption generation [153], and sentiment analysis [154]. Such biases can be mitigated by creating augmented dataset that counters gender bias in the original training dataset [154], or add constrains in model training to enforce gender neutral prediction [155].

6.3 Detect Gender Stereotype: A Lexicon Approach

In this section, we introduce the lexicon approach for detecting gender stereotypes in articles. The methodology is based on a gender stereotype lexicon, which use machine learning to derive a *gender score* for each adjective and verb. The score reflects the word's masculinity or femininity, learned through a set of ground truth data we obtained using user survey, in which we let users rate how easily they associate words to gender. Finally, we use the lexicon to compute gender score of an article which judges how consistent or contradictory the article is with gender stereotypes.

6.3.1 Data Collection

Candidate Word Selection

We aims at extracting a large set of words that are potentially related to gender stereotypes. Here we restrict our selection to *verbs* and *adjectives*, as stereotypes often exist in people's behavior and how they are described [21].

The candidate words are extracted from Wikipedia Datadump[1]. We choose Wikipedia data because it is large and diverse, so it is likely to include most commonly used English words. We downloaded the lasted version of Wikipedia text on March 4th, 2019, removing all images and links. In total, our dataset includes 5,817,125 documents, 42,653,358 paragraphs and 2,076,621,930 words.

To extract verbs and adjectives that characterize human, we analyze the word dependencies in each sentence using parse tree. For verbs, we extract "subject-verb" relationship in each sentence, where the subject is human-related words like "he", "she", "man", "women" *etc.* For example, in the sentence "He ran away from her," the subject is "he" and the verb is "ran," so the word "ran" is extracted. We lemmatized the words so the past tense "ran" become the lemma "run". For adjectives, we consider two different types: *predicate* and *attribute*. Adjectives are predicate when they are connected to their subject words by a verb, usually "be", *e.g.,* "he is *handsome*." Attribute adjectives are used as modifier before the subject, as in "a *handsome* man." In both cases, we keep the adjective if it is used on a human-related word. The parse tree tool we applied is implemented using SpaCy [2].

We then filter out words that cannot be found via the Oxford Dictionary API [3]. The removed words are mostly non-English words, non-existing words, and words with the

[1]https://dumps.wikimedia.org/
[2]https://spacy.io/
[3]https://developer.oxforddictionaries.com/documentation

wrong part-of-speech (*e.g.,* a word extracted as an adjective but is only used as a noun). After our dictionary check, 6,178 verbs and 4,424 adjectives remain in our dataset.

We select the most commonly used 1,000 verbs and 1,000 adjectives to be labeled through user survey, serving as ground truth for our gender score computation models. Three authors manually checked all the selected words to make sure that they are suitable candidates. We removed words that are related to race, country or religion, and words that depend strongly on context (*e.g.,* "next", "final").

Survey Design

Following previous studies [142, 150], our survey asks the participants to rate the extent to which they associate each word shown with a typical man or woman. Specifically, the participants are shown a list of words, and asked to evaluate the statement "I feel that _____ is commonly associated with the characterization of a typical man" or "of a typical woman" in 7-point Likert Scale, from "strongly disagree" to "strongly agree." They can also select "I don't understand the word." Each participant rate 50 adjectives and 50 verbs "of a typical man" (*male rating*) and another 50 adjectives and 50 verbs "of a typical woman" (*female rating*). We collect their demographics information in the end. The whole survey takes about 15 minutes and the participants are compensated $3.

To make sure that the participants pay attention during the rating, we also randomly insert 4 words that do not exist in English. The participant are expected to select "I don't understand the word." for all 4 words. We also include another quality control question when collecting demographics information that asks the participants to choose A and D. In our data analysis, we removed all the responses that failed these quality check questions. We attach a full survey script in Appendix A.3.

We recruited our participants on Amazon Mechanical Turk [4]. Each participant can

[4] https://www.mturk.com/

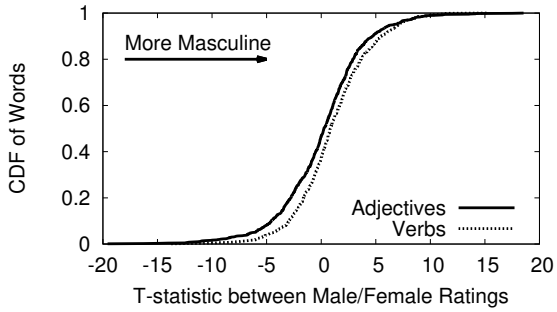


Figure 6.1: CDF of T-statistic between male ratings and female ratings. A higher number indicate people more easily associate the word to men than to women.

answer our survey for up to 10 times, but they will get different words each time. We collected a total of 1097 response sessions (HITs), among which 619 are from male participants, 476 are from female participants, and 2 are not willing to disclose their gender. Over 99% of the words have more than 50 male ratings and 50 female ratings.

6.3.2 Data Analysis

Gender Score Calculation

The ground truth score of each word is measured by the difference between the male ratings and the female ratings. For example, if a word is perceived as strongly associated with man but not associated with woman, then the word should carry strong masculine stereotype.

Specifically, we use the T-statistic in a two sampled T-test to measure the difference between male ratings and the female ratings of a word. The T-statistic reflects the extent to which the average value differs across samples. Like other statistical tests, the T-statistics also maps to a p-value which indicates how likely the average value of the two samples are identical. A small p-value means that statistical significant different between samples [156].

(a) Adjectives (b) Verbs

Figure 6.2: Adjectives and verbs that are significantly masculine and feminine. Words in green are feminine words and words in red are masculine words. Larger font size indicates higher masculinity/femininity (larger T-statistic magnitude).

We plot a the distribution of T-statistic of each word in Figure 6.1. Most words score around 0, which indicates gender neutrality. Figure 6.2 shows examples of words with different T-statistics. We attach a full word list in Appendix A.4. Except a few words related to appearance (*e.g.,* hairy, beautiful), our stereotype words are consistent with recent works demonstrating that stereotypically men are strong, active and violent, and women are weak, emotional and dependent [21].

Reliability of User Responses

We perform two tests to examine the reliability of our responses. The first test is *split-half reliability* [157]. Split-half reliability aims at measuring how likely the data collection is *reproducible*. To do so, we randomly split all our participants in to two equal-sized halves, and then calculate two sets of T-statistics for each word independently using the two halves. Finally, we calculate the Pearson Correlation between the T-statistics of the

two sets, getting 0.85 for adjectives and 0.82 for verbs, which indicate that repeating the data collection process will not significantly change the result.

In our second test, we determine to what extent responses from male and female participants agree with each other. Similar to the calculation of split-half reliability, we split our responses by the gender of the participant, calculate two sets of T-statistics for each word independently using the two splits. From the two sets, we get a correlation of 0.82 for adjectives and 0.80 for verbs, which is close to the correlations in the split-half reliability. The results show that no significant difference exists between responses of male and female participants. Thus, in the following analysis, all our scores are calculated by mixing all responses together.

6.3.3 Gender Stereotype Lexicon

Using the 2,000 words as ground truth, we seek to compute and validate the gender scores of all the verbs and adjectives in our dataset. We consider approaches: unsupervised and supervised. Unsupervised methods assign score to words based on pre-trained word embedding model or large scale text corpus. They are fully automated and do not rely on human label. Supervised methods train a regression model based on scores of the sampled words. We find them to be much more accurate than the unsupervised methods.

Unsupervised Methods

Our unsupervised methods are inspired by previous works that calculate gender information of words. Although they are not fully optimized for human-defined gender stereotype, they are similar to gender stereotypes in some extent. Our methods are the following:

- **Odds ratio.** Odds ratio calculates how likely a verb or an adjective is used to characterize a man rather than a woman. If a word is more likely to be used on a man, it may indicate masculinity of the word. Specifically, given a word, odds ratio is calculated as:

$$\frac{\#\ this\ word\ on\ man / \#\ this\ word\ on\ womam}{\#\ other\ words\ on\ man / \#\ other\ words\ on\ woman} \tag{6.1}$$

  . Here, "#" denotes "number of times." Odds ratio reflects gender stereotype in large language corpus [21]. We calculate the odds ratio using the Wikipedia data.

- **Distance to gender specific words.** It has been shown that word embeddings contains human-like stereotypes [158]. Such stereotype can be captured by calculating word distance to gender specific words. Specifically, given a word, we calculate its average distance to a set of male specific words (*e.g.,* "he", "man") and its average distance to a set of female specific words (*e.g.,* "she", "woman"), and then calculate difference between the two distances. We test 3 most commonly used word embeddings: word2vec [159], GloVe [160], and FastText [161]. We do not train our own word embeddings, but apply widely used pre-trained models for all of the 3 embedding methods: word2vec from Google News [5], GloVe from 6 billion token Wikipedia [6], and FastText from English Wikipedia [7].

- **Projection on gender direction.** One way to reduce gender stereotype in word embedding is to extract a gender direction and remove the vector projection on the direction [137]. Here, the gender direction is the direction parallel to $\overrightarrow{she} - \overrightarrow{he}$ or $\overrightarrow{woman} - \overrightarrow{man}$. For each word, we take its projection on the gender direction as the gender stereotype score of the word.

[5] https://code.google.com/archive/p/word2vec/
[6] https://nlp.stanford.edu/projects/glove/
[7] https://fasttext.cc/docs/en/pretrained-vectors.html

Method	Adjective	Verb
Odds ratio	0.09	0.29
Distance + word2vec	0.44	0.37
Distance + GloVe	0.47	0.41
Distance + FastText	0.47	0.41
Gender direction	0.40	0.33
Gender dimension	0.20	0.08

Table 6.1: Pearson Correlation of scores calculated by unsupervised methods and ground truth on all 1,000 verbs and 1,000 adjectives.

- **Values on gender dimensions.** Another way to reduce gender stereotype in embedding is to encode gender information in a reserved dimension during training [155]. Here, we use the magnitude of the gender dimension as a way to quantify gender stereotype of a words. We use pre-trained word embedding provided by the paper authors.

For evaluation, we calculate the scores for each word in our ground truth data, and examine the Pearson Correlation between the calculated score and the ground truth *gender scores*. As shown in table 6.1, scores calculated by all of these methods are positively correlated with human defined gender stereotype, but the magnitude of the correlation is no larger than 0.5.

Supervised Methods

Our supervised methods train machine learning models that predict the gender score of each word based on pre-trained word vectors. Compare to unsupervised methods, supervised methods generate more accurate scores on our ground truth dataset.

We use word embeddings of each word as features. Similar to unsupervised methods, we use pre-trained word2vec, GloVe and FastText as model input. Given that the size of our dataset is only a thousand words, we do not apply deep neural network which usually requires large amount of training data. Instead, we apply standard machine learning models: Support Vector Machine (SVM) and Linear Regression (LR).

Model	Feature	Corr. on Verb	Corr. on Adjectives
LR	word2vec	0.57	0.63
SVM	word2vec	0.57	0.62
LR	GloVe	0.52	0.53
SVM	GloVe	0.55	0.58
LR	FastText	0.45	0.52
SVM	FastText	0.53	0.58

Table 6.2: Pearson Correlation of scores calculated by supervised methods and ground truth.

For evaluation, we use 80% of words for training, and calculate Pearson Correlation between prediction and the ground truth for the remaining 20% of the words. We repeat the test 100 times, and calculate the average correlation. We list our results in Table 6.2. We can see that the correlations for the learned methods are higher than correlations of the unsupervised methods. Linear Regression with word2vec as feature has the highest correlation. In the following sections, we extend gender score to all the words in our data using scores derived from the Linear Regression model using word2vec as features.

6.3.4 Gender Score of Articles

Finally, we introduce our method that detect gender stereotypes of an article using our lexicon. Similar to previous work [25], our lexicon approach assigns a stereotype score to an article based on the word usage. We use the gender score of words in our lexicon to determine the score of an article: we extract all the verbs and adjectives in the article, and add scores of these words together.

In the next section, we will formulate the gender stereotype detection task as two classification tasks. The gender score of an articles can be transferred into a label to fit the classification settings. Specifically, if the total score is positive, the paragraph is labeled as consistent with masculine stereotype or contradictory to feminine stereotype, otherwise the label is consistent with feminine stereotype or contradictory to masculine

stereotype.

In section 6.5, we will apply the lexicon approach to the classification tasks, and compare the performance to an end-to-end approach. We test the lexicon approach on two lexicons: using 2,000 selected words in our survey, and after we extend the score to all the words.

6.4 Detect Gender Stereotype: An End-to-End Approach

In this section, we introduce the end-to-end approach. Our end-to-end approach is based BERT [162], a language representation tool that converts articles into vectors. We formulate the gender stereotype detect task as two binary classification tasks, build a dataset that ask users to provide articles that are consistent with or contradictory to common gender stereotypes, and train the BERT model to generate classification labels.

6.4.1 Task Definition

We formulate the gender stereotype detection problem as *two binary classification problems*: determining whether the description of a man is consistent with masculine stereotype, and determining whether the description of a woman is consistent with feminine stereotype.

The input of both tasks is an article. For the masculine stereotype task, the article is a piece of description of a man. For the feminine stereotype task, the article is a piece of description of a woman. The labels are “consistent” (positive) or “contradict” (negative) for both classification tasks.

	Consistent	Contradict
Masculine	974	975
Feminine	973	974

Table 6.3: Number of articles in different combination of requirements.

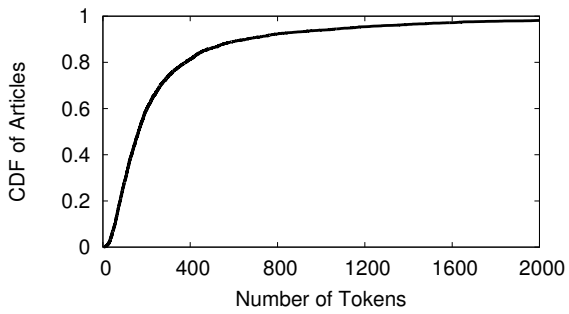


Figure 6.3: CDF of number of tokens in articles.

6.4.2 Data Collection

To collect the articles for our prediction tasks, we perform a survey study. Our survey ask users to search the Internet, and copy paste articles (or a few paragraphs of an article) that meet the following requirements: it describes a man (or woman), and the description is consistent with (or contradictory to) common gender stereotypes. We also ask them to briefly state the reason for choosing each article. Each user is asked to provide 4 articles, with 4 different combinations of requirement (man or woman, consistent or contradict). The entire survey takes about 25 minutes and the participants are compensated $3. We attach the full survey script in Appendix A.5.

We recruited our participants on Prolific[8]. Prolific is a crowdsourcing service aiming at providing high quality data that empowers research. We collected a total of 980 responses, 508 of which are from male participants, 457 are from female participants, and 15 are from people unwilling to disclose their gender. We filtered out 24 articles that do not contain any pronouns, named entities or gender specific words, indicating that

[8]https://prolific.ac/

Domain	Number	Domain	Number	Domain	Number
wikipedia.org	366	dailymail.co.uk	52	time.com	41
nytimes.com	149	forbes.com	50	foxnews.com	41
theguardian.com	73	people.com	49	cnbc.com	36
cnn.com	68	huffpost.com	44	biography.com	34
npr.org	55	washingtonpost.com	43	usatoday.com	30

Table 6.4: Top domains and number of articles from each domain.

	Consistent	Contradict
Masculine	championship, ceo, president, gun, mountain, top, businessman, league, fight, win, football, service, player	gay, makeup, gender, singer, fashion, comfortable, mom, caregiver, cosmetic, stay, youtube, openly, feel, kid, identity
Feminine	cook, home, child, beauty, care, clean, instagram, housewife, daughter, baby, mother, prince, fashion, family	champion, sport, field, fighter, athlete, history, martial, force, hole, fight, technology, team, institute, gender

Table 6.5: Top keywords that distinguish consistent with stereotype and contradict stereotype.

these articles are not likely to be descriptions of people. Table 6.3 shows the number of articles in each category. The length of the articles varies a lot, as shown in Figure 6.3. When looking at the sources of the articles, most of the articles are from biography page (*e.g.,* Wikipedia), or from news (*e.g.,* New York Times). We list the most frequently used domains in Table 6.4.

To understand the content of the articles, we extract top key words in each category using Chi-square statistics [163]. Chi-square statistics measure how strongly a word can be used to distinguish articles in different categories, *i.e.,* the consistent and the contradictory. We calculate Chi-square statistics for masculine stereotype and feminine stereotype separately, and list the top keywords in Table 6.5. The keywords show that our survey participants choose competitive man and caring women to exemplify gender stereotypes. Meanwhile, there are some similarity between man who contradicts stereotype and woman who are consistent with stereotype (and vice versa).

We use the articles that describe men for the masculine stereotype task, and use articles that describe women for the feminine stereotype task. We randomly split out data into chunks of 8:1:1 for training:validation:testing. The training and validation data are used to build the end-to-end model. The test data is used to compare the performance of between the lexicon approach and the end-to-end approach.

6.4.3 Classification Model

Finally, we introduce our classification model. Our model is trained on top of BERT [162], an unsupervised language representation tool pre-trained on large language corpus. BERT uses a fully connected transformer structure. It is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. A significant advantage of BERT is that it can be fine-tuned with one additional output layer to create a state-of-the-art model for a wide range of tasks, including document classification. So we use BERT as a representative of state-of-the-art deep neural network model.

We use our training data to fine-tune the model, and use the validation set to identify the optimal hyper-parameters, which is 2e-5 learning rate for 3 epochs.

6.5 Performance Evaluation

In this section, we apply both the lexicon approach and the end-to-end approach on the test data to classify if articles are consitent with or contradict gender stereotypes. Our results show that the end-to-end approach significantly outperforms the lexicon approach by overcoming a few fundamental problems of the lexicon approach. The end-to-end approach does not require a large sample of training data to perform well. Our end-to-end approach can be adjusted to detect gender bias in job advertisements that

	Accuracy (Masculine)	AUC (Masculine)	Accuracy (Feminine)	AUC (Feminine)
BERT	0.81	0.85	0.76	0.85
Lexicon (Extended words)	0.68	0.71	0.68	0.70
Lexicon (Selected words)	0.62	0.65	0.67	0.68

Table 6.6: Accuracy and AUC of lexicon approach and end-to-end approach.

outperforms the industry state-of-the-art.

6.5.1 Prediction Performance

We apply algorithms from both the lexicon approach and the end-to-end approach to the test set, and calcuate prediction accuracies and AUCs. The performance is shown in table 6.6. The results show that the end-to-end approach significantly outperforms the lexicon approach, both in terms of accuracy and AUC. For lexicon approach, using extended word scores do help increase the performance.

6.5.2 Understanding the Lexicon Approach

To understand the lexicon approach generates unsatisfying prediction results, we manually examine the *all* the incorrect prediction the lexicon approach makes in the test set, and summarize the possible reasons behind the misclassification. We summarize the reasons and provide examples in Table 6.7. For each reason, we also calculate how many times the lexicon approach make mistakes (the "Lexicon Wrong" columns) and how many times the end-to-end approach make wrong prediction among these cases(the "and E-to-E Wrong" column). The detailed explanations are as follows:

- *Lexicon Coverage* Our lexicon only covers adjectives and verbs, and gender stereotypes can be expressed by words outside of our lexicon. For example, "PhD" could be a word that reflect masculine stereotype.

Reason	Lexicon Wrong	and E-to-E Wrong	Example
Lexicon Coverage	6	0	The first woman I invited to co-author a **publication** was in 2015, four years after completing my **PhD**.
Phrase	8	0	... who **paints his fingernails**, **braids his hair** and **poses for gay magazines**...
Non-human	5	0	Katie Bouman has already worked on looking around corners by analyzing **tiny** shadows ...
Word sense	4	1	The Latin origin of "integrity" means whole, and when it comes to being a good guy, wholesome is **sexy**.
Consistent and contradict	22	5	Even as I regularly work out and **lift weights**, I am a rather **fragile** excuse for a woman, constantly getting sick...
Multiple people	10	2	**My wife** had more earning potential and so I volunteered to concentrate on family and home.
Subtle stereotype	25	14	*Trump turned his wife's birthday into an official meeting, which contradicts masculine stereotype that men should respect his wife's birthday.*
Insufficient information	19	6	*American actor Peter Dinklage contradicts masculine stereotype because he is a dwarf, which is not discussed.*
Data noise	24	15	*Random response or fail to meet task requirement.*

Table 6.7: Reasons for lexicon approach making wrong classification. The "Lexicon Wrong" column is the number of cases when lexicon approach make wrong prediction, and the "and E-to-E Wrong" column is the number of cases the end-to-end approach are wrong among these cases. Bold words are words that are closely related to the reason. Italic words are not exact content from our data, but explanations.

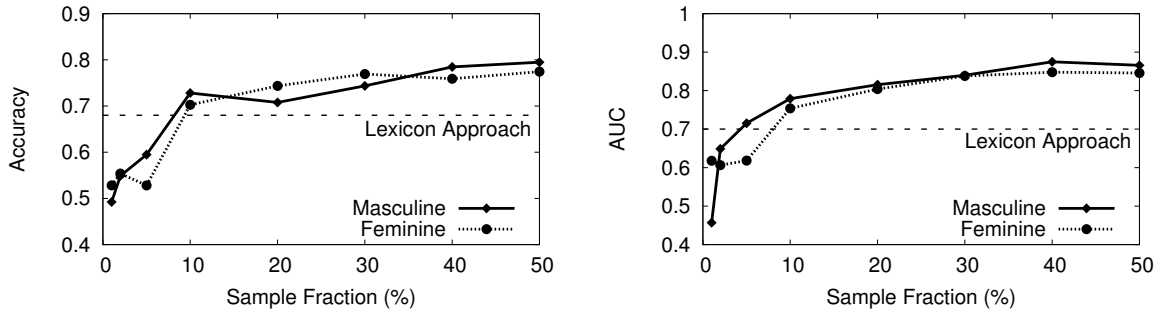
- *Phrase* The stereotype is expressed by a multiple-word description, which can not be captured by single words.
- *Non-human* The word that contains strong gender stereotype is used on a non-human object. For example, in "tiny shadows", "tiny" is labeled as a feminine word but "tiny shadows" does not indicate femininity.
- *Word sense* A word could have multiple senses, while only one sense suggest gender stereotype. For, "sexy" usually have a feminine meaning of "sexually attractive", but

it could also mean "healthy" which is gender neutral.

- *Consistent and contradict* The article contains description of a person who has some characteristics that are consistent with stereotype and some other characteristics that contradict. Although the article have some focus on one side, the lexicon approach can not identify the true focus by word count.
- *Multiple people* The article describes a few different people, usually one people as the main character while the others are supporting characters. The lexicon approach can not catch the descriptions of the correct person.
- *Subtle stereotype* A few users may have different understanding of stereotypes. For example, a user puts "respect wife's birthday" as a masculine stereotype.
- *Insufficient information* The stereotype is not in the text description, but in general knowledge. For example, an article about American actor Peter Dinklage is labeled as contradict with masculine stereotype because he is a dwarf, but the fact can not be found in the article.
- *Data noise* These are the low quality response including cases when the users provide responses that do not fit our task requirement (*e.g.,* the paragraph is not description of a person, article is consistent with stereotype when we ask for contradiction).

6.5.3 Data Size for End-to-end Approach

End-to-end learning approaches often require a large amount of training data. Here, we seek to quantify how many training data is needed to outperform the lexicon approach. We vary the size of the training data, train the BERT fine-tuning model, and evaluate the performance on the same test set. The results are shown in Figure 6.4. We have two observations. First, the performance plateaus when the training data size reaches 40% of our full training data, beyond which further increase in training data size yields no



(a) Accuracy (b) AUC

Figure 6.4: End-to-end approach performance with different training data size, compare to lexicon approach (similar number for masculine task and feminine task).

	Textio	Unitive	BERT fine-tune
% of woman	0.59	0.54	0.78
Attractiveness to female applicants	0.64	0.54	0.81

Table 6.8: Pearson correlation between user response and gender bias scores.

performance gain. Second, the end-to-end approach can outperform the lexicon approach even when the training data is only 10% of current size, which is about 150 articles. This indicates that the end-to-end approach does not need a large corpus to learn typical gender stereotypes.

6.5.4 Application: Gender Bias in Job Postings

Finally, we apply our end-to-end approach to detect gender biased language in job postings, and compare the performance to the industry state-of-the-art. The data comes from a survey study in a previous work [25]. In the survey, a sample of 30 job postings was shown to the participants. They were asked to read the job postings, and answer a few questions about the extend to which the job posting is gender biased. Here, we use their answers on 2 questions to quantify the gender bias in the job postings: 1) the percentage of women in the position (0%-100%) and 2) how likely the job posting is attracting more female applications or male applicants (7-point Likert Scale, from attracting mostly male

applicants to attracting mostly female applicants).

To convert our male and female stereotype detection models into a single gender bias indicator, we take the job posting text and use it as input in both the masculine stereotype classifier and feminine stereotype classifier. We take the probability output from both classifiers, and calculate the difference in the two probabilities. For example, if a job posting is predicted as 90% likely to be consistent with masculine stereotype and 60% likely to be consistent with feminine stereotype, the score of the job posting is 0.3, which indicates masculine bias.

We calculate the score for all the job advertisements, along with two state-of-the-art services used in the previous work: *Textio* and *Unitive*, both of which are specifically designed for detecting gender bias in job posting using lexicon approach [25]. Table 6.8 show the correlation between the scores and user response. This shows that although the models in our end-to-end approach is not specifically trained for job advertisements, they still outperform the best lexicon approaches designed for this task.

6.6 Conclusion

Our work systematically compares two approaches for detecting gender stereotype from text. We collect the first labeled dataset that maps articles to gender stereotype labels. We develop algorithms using both a lexicon approach and an end-to-end approach for detecting gender stereotypes. We apply both approaches on our dataset, and show that end-to-end approach significantly outperforms lexicon approach, by overcoming a few fundamental limitations of the lexicon approach. Our experiments show that the end-to-end approach do not require a large training corpus to perform reasonably well, and it can be directly applied to real applications such as detecting gender bias in job postings. We hope that by highlighting issues with existing stereotype detection approaches, we

encourage wider use of the end-to-end approach and more studies on improving our models.

Our work do have limitations. First, we formulate the task as binary classifications, where a neutral label is not considered. The overall good performance may come from a oversimplified classification settings. Second, we only collect a relatively small dataset with a few thousands of training samples, so it is still unclear whether the performance can be further improved if the training data is orders of magnitude larger. Third, our end-to-end approach comes from fine-tuning the current state-of-the-art. It is possible that a more task-specific model can generate a better performance. We hope that future works can better address these limitations

Chapter 7

Conclusion

In this dissertation, we have investigated three different types of cognitive bias, and the outcome of these biases when people form group and interact with each other. Our methodology aims that measuring biases in the wild and add evidences to existing theories. We study irrational user behavior in financial domain in Chapter 3 and Chapter 4. Then we analyze gender bias in different real world scenarios in Chapter 5 and Chapter 6.

In this Chapter, we summarize the contribution of our work, and discuss potential future directions in this research area.

7.1 Summary

Cognitive biases have been an important topic in psychology over the past 60 years. In a long period of time, analysis of cognitive biases are based on lab experiment that observe human behavior in a controlled environment. With the development of the Internet, more and more user actions have been moved online, which generates rich user behavior data that supports large-scale analysis of cognitive biases in the wild.

In Chapter 3, we study the quality of user generated content of investment discussion

boards. We find that user sentiment do not reflect real world stock price change, and sentiments of discussions are controlled by a small group of users. Our results suggest that discussions on these online forums are likely to be echo-chamber, and they might play a significant role in introducing noise trading by individual investors.

In Chapter 4, we look for evidences that investors herd when there is a lack of high quality financial information in immature stock market. We compare US and Chinese stock market in terms of adoption of technical analysis. Our user study shows that a majority of Chinese investors use technical analysis, mainly because it is recommended by media and friends. Our analysis of two decades stock price data shows that technical analysis tools in China are significantly more accurate than in the US, and the accuracy increase with the popularity of the tool. Our results suggest that individual investors in Chinese market do herd, which produce synchronized signals in stock price change.

In Chapter 5, we study the role of gender biased language in introducing and exacerbating gender inequality in the job market. Our analysis of 10-year job postings on LinkedIn shows that although the job market is getting more and more gender neutral over time, significant bias still remains in individual positions. Our user study demonstrates that although language do slightly affect user application decision, people still hold strong gender stereotypes. Our work suggests that we need to address inherent gender bias in applicants themselves to improve gender neutrality in the job market.

In Chapter 6, we systematically compare the performance of a lexicon approach and an end-to-end approach in gender stereotype detection. We build a gender stereotype lexicon, and collect articles with human labeled gender stereotype information. We apply both approach on our data, and shows the end-to-end based approach significantly outperforms the lexicon approach by mitigating fundamental limitations of lexicon approach. Our work suggests that in the future the end-to-end approach may replace widely used lexicon approach for detecting gender bias in text.

7.2 Future Work

7.2.1 Increase Awareness of Unconscious Bias

Often cognitive bias functions in a way that is out of users conscious awareness. So in order to fight the negative consequences caused by the cognitive bias, it would be helpful to increase user awareness of the bias.

One future direction that directly follows the studies in this dissertation is to take bias into consideration during user interface design. For example, we have shown that online communities such as investment discussion boards are vulnerable to echo-chamber effect. User sentiments are monotonic and isolated from the market change. To mitigate this effect, we can leverage interface design to encourage diverse point of views, and incorporate new information from the wild. These could be potentially achieved by embedding news information to the discussion sentiment, and rewarding new ideas and opinions. As another example, we have shown that language in job advertisements may affect decision of potential applicants. To mitigate this effect, we can add hints on job listings on the recruiting websites that show potential language bias in the job listings.

Another direction is to develop training programs that teach people how biases can lead to negative consequences. For example, in online forums when user reply indicate possible exercise of confirmation bias, a training message about potential negative consequence pops up. Similar ideas have been widely used in computer security like preventing phishing (*e.g.,* send user phishing emails and launch training if users falls), but few has been applied to promote awareness of biases.

7.2.2 Gender Fair Language

Content is only one aspect of language. *Grammatical structure* of language may also contribute to perceived gender bias. Language used on male person and female person are *asymmetric*. For example, in English, nouns referring to female persons are composed by adding suffix to corresponding nouns referring to male persons (*e.g.,* hero vs. heroine, actor vs. actress) [164]. Also, masculine forms are often used as generic to refer a group or unspecified gender (*e.g.,* "actors" refers to a group of actors and actresses).[165]. In some other cases, people tend to compare groups by mentioning the higher status group first (*e.g.,* , "compare to men, women are ...") [166].

By carefully choose the right language to use, such language asymmetry can be avoided (*e.g.,* use "police officer" instead of "policeman"). This way of using language is called gender-fair language [139]. Gender-fair language is designed for fighting against stereotypes and discriminations. As a follow-up of our work, we could develop tools that can automatically covert any normal language into gender-fair language. We hope that adoption of a language auto-correction tool could significantly reduce gender biased encoded in natural language.

Appendix A

Appendix

A.1 Survey for Stock Investment Strategies

1. How many years of experience do you have investing in the stock market?

   - I do not invest
   - 0-1 year
   - 1-5 years
   - 6-10 years
   - 11-20 years
   - 20-30 years
   - 30-40 years
   - >40 years

2. How would you describe your degree of risk tolerance in the stock market?

   - **Low**: I do not want to lose money, even if it means my returns are relatively small.
   - **Moderate**: I am willing to accept the occasional loss as long as my money is in sound, high-quality investments that can be expected to grow over time.
   - **High**: I am willing to take substantial risk in pursuit of significantly higher returns.

3. There are two basic investment strategies used to predict stock prices: technical analysis and fundamental analysis.
   Technical analysis examines a stock's past prices and volumes. Charts and indicators are used to identify price patterns that suggest future price movements. In this strategy, price is the key predication tool.
   Fundamental analysis examines the underlying health of a company, and thereby the true price of its stock, by looking at its core facts (e.g., income statements, earnings releases, balance sheets, company news). It also takes general economic factors (e.g., industry conditions, the unemployment rate, monetary policy) into consideration. The key is to estimate a company's intrinsic value.
   Please rate how frequently you consult each of these investment strategies (Most frequently being "Every day," and least frequently being "Never").

	Every day	Several times per week	Once a week	Once a month	Less than once month	Never
Technical	○	○	○	○	○	○
Fundamental	○	○	○	○	○	○

4. If you check with technical analysis to inform your investment decisions, how often do you use the following technical indicators?
(Please rate each of the indicators from use it "Every day" to "Never/Not Applicable". If you do not use technical analysis, please select "Never/Not applicable" for each row.)

	Every day	Several times per week	Once a week	Once a month	Less than once month	Never/Not applicable
MA	○	○	○	○	○	○
MACD	○	○	○	○	○	○
RSI	○	○	○	○	○	○
STO	○	○	○	○	○	○
KDJ	○	○	○	○	○	○
MFI	○	○	○	○	○	○
CMF	○	○	○	○	○	○
CCI	○	○	○	○	○	○
TRIX	○	○	○	○	○	○

If you use other indicators, please specify the name and frequency (optional)

5. If you check with technical analysis to inform your investment decisions, what is/are your primary reason(s) for doing so? Select all that apply.

   - ☐ I do not use technical analysis.
   - ☐ Technical analysis is provided by my stock software.
   - ☐ Technical analysis is recommended by my friends.
   - ☐ I've learned technical analysis from traditional media, e.g., broadcast on TV/radio.
   - ☐ I've learned technical analysis from social media, e.g., blogs and stock forums.
   - ☐ I consult technical analysis because fundamental analysis is not accurate enough.
   - ☐ Technical analysis is faster to learn than fundamental analysis.
   - ☐ Technical analysis is easier to use than fundamental analysis.
   - ☐ Technical analysis needs no other information other than stock history.
   - ☐ For other reasons, please specify (optional)

6. If reliable information about companies and the stock market (e.g., investment research reports) were more openly available (i.e., free to access, easier to find, etc.), how would this affect your investment strategy?

   - I would use more fundamental analysis and less technical analysis.

- I would keep my current investment strategy.
- I would use less fundamental analysis and more technical analysis.
- I do not know.
- For other impacts, please specify (optional)

7. What additional resource(s), if given to you, would most help you make better informed stock investment decisions? Select all that apply.

   - ☐ More complete investment/stock market news.
   - ☐ Local investment communities.
   - ☐ Personal investment consultants.
   - ☐ Subscription to investment/financial newspapers or journals, e.g., Wall Street Journal, Barron's, Financial Times.
   - ☐ Academic research on stock markets and investment.
   - ☐ Improved stock analysis software.
   - ☐ For other resources, please specify (optional)

8. If you had to choose one, would you prefer to invest in the U.S or the Chinese stock market?

   - I prefer the U.S. market.
   - I prefer the Chinese market.
   - I do not know.

9. What is your gender?

   - Female
   - Male
   - Other

10. What is your age?

    - <20
    - 20 - 30
    - 31 - 40
    - 41 - 50
    - 51 - 60
    - 61 - 70
    - 71 - 80
    - 81 - 90
    - 91 - 100
    - >100

11. How long have you lived in U.S.?

- <1 year
- 1-5 years
- 5-10 years
- 11-20 years
- 21-30 years
- 31-40 years
- 41-50 years
- 51-60 years
- 61-70 years
- >70 years

12. What is your annual (approximate) household income?

- $0 - $25,000
- $25,001 - $50,000
- $50,001 - $100,000
- $100,001 - $150,000
- $150,001 - $200,000
- $200,001 - $250,000
- $250,001 - $300,000
- $300,001 - $400,000
- $400,001 - $600,000
- $600,001 - $800,000
- $800,001 - $1,000,000
- >$1,000,000

A.2 Survey for Evaluating Job Advertisements

Main Survey Questions

Please read the following job description and answer the questions below: *Show job advertisement here*

1. If you were fully qualified to apply for a job like this, how likely is it that you would apply for this particular position?

   - 1 (= I would definitely not apply)
   - 2
   - 3
   - 4
   - 5 (= I would definitely apply)

   Please tell us the reasons for your answer: _____

2. By looking at the job description, what would you think to be the percentage of women currently working in this type of position?

   - 0 (= 0%)
   - 1
   - 2
   - 3
   - 4
   - 5
   - 6
   - 7
   - 8
   - 9

- 10 (= 100%)

Please explain why you selected the answer you did: _____

3. While reading the job description, to what extent did you feel that the advertisement would attract more male or more female applicants?

- 1 (=This job ad will likely attract mostly male applicants)
- 2
- 3 (=This job ad will attract male and female applicants equally)
- 4
- 5 (=This job ad will likely attract mostly female applicants)

Please tell us the reason for your answer: _____

4. Please mark any words or sentences that you do not feel comfortable with.

Demographics

1. What is your gender?

- Female
- Male
- Other

2. What is your age?

- <20
- 20 - 30
- 31 - 40
- 41 - 50
- 51 - 60
- > 60

3. What is your ethnicity?

- Black or African American
- Asian
- Hispanic or Latino
- Pacific Islander
- White
- Other

4. What is the highest level of education you completed?

- Did not complete high school
- High school/GED
- College
- Bachelor's degree
- Master's degree
- Advanced graduate work or Ph.D.

5. Which of the following best describes your current occupation?

   - Student
   - Full-time employment
   - Part-time employment
   - Self-employed
   - Unemployed
   - Retired

6. If you answered student to (5), what is your major?

7. If you answered employed to (5), what is the job industry that best describes your current job?

A.3 Survey for Word and Gender Association

Task 1 Instruction

On the following pages, you will be shown lists of adjectives. We are trying to determine **the extent to which people think each adjective is associated with men or with women.**

Your task is to evaluate the extent to which you associate each adjective shown with a typical man or woman.

In other words, if you think the adjective it is typically used to characterize a man (or a woman), then it is considered to be associated with a man (or a woman).

You will evaluate each adjective for its association with a man or woman, but not both.

Just to make sure you understand the task, here is an example: If the word shown is "square" and you strongly disagree that this word is typically associated with a man, then you would answer 1 on the scale below.

Task 1 begins now:

I feel that _____ is commonly associated with the characterization of a typical MAN.

Show a list of 50 adjectives, rate agreement in 7 point Likert Scale

I feel that _____ is commonly associated with the characterization of a typical WOMAN.

Show a list of 50 adjectives, rate agreement in 7 point Likert Scale

Task 2 Instruction

Your task is to evaluate**the extent to which you associate eachVERB shown with a typicalman or woman.** This is similar to Task 1.

In other words, if you think the verb is typically used to characterize a man (or a woman), then it is considered to be associated with a man (or a woman)

Please evaluate the verb as though a man (or a woman) is the **subject** (i.e. the person doing the action). Just to make sure you understand the task, here is an example: If the verb shown is ran, imagine the following sentence: "The man ran away from the fox."

We would like you to evaluate how strongly you associate the verb with respect to the **man**, not the fox.

You will evaluate each verb for its association with a man or woman, but not both.

Task 2 begins now:

I feel that _____ is commonly associated with the characterization of a typical MAN.

Show a list of 50 verbs, rate agreement in 7 point Likert Scale

I feel that _____ is commonly associated with the characterization of a typical WOMAN.

Show a list of 50 verbs, rate agreement in 7 point Likert Scale

Demographics

1. How old are you?
   - 18 - 19
   - 20 - 30
   - 31 - 40
   - 41 - 50
   - 51 - 60
   - 60+
2. What is your gender?
   - Female
   - Male
   - Other
3. What is your ethnicity?
   - Black or African American
   - Asian
   - Hispanic or Latino
   - Pacific Islander
   - White
   - Other
4. What is the highest level of education you completed?
   - Less than a high school diploma
   - High school or equivalent (GED)
   - Some college (no degree)
   - Associate's degree (e.g., AA, AS)
   - Bachelor's degree (e.g., BA, BS)
   - Master's degree (e.g. MA, MS)
   - Advanced graduate work or Ph.D. (e.g., PhD, MD, DDS)
5. For quality control, please choose A and D for this question (please choose both):
   - A
   - B
   - C
   - D
   - E

6. Which of the following best describes your current occupation?

- Student
- Full-time employment
- Part-time employment
- Self-employed
- Unemployed, looking for work
- Unemployed, not looking for work
- Retired

A.4 Words in Different Gender Score Range

Adjectives, Strongly Masculine ($t > 3.50, p < 0.001$)

abrasive, abusive, adamant, adulterous, adventurous, aggressive, alcoholic, ambitious, armed, arrogant, assertive, athletic, authoritative, autonomous, bald, bearded, bent, big, blunt, bold, brave, brutal, charismatic, cheap, chief, cocky, combative, competitive, conservative, constructive, convincing, corrupt, courageous, criminal, crooked, cruel, dangerous, daring, dark, decisive, defiant, destructive, direct, dirty, disrespectful, dominant, drunk, drunken, enraged, expert, fearless, firm, focused, forceful, formidable, frank, giant, greedy, hairy, handsome, handy, hard, harsh, headstrong, heavy, heroic, high, historic, homophobic, horrible, hostile, huge, ignorant, immature, immune, impotent, independent, indifferent, industrious, influential, innovative, insensitive, intense, intolerant, invincible, invulnerable, large, leading, loud, major, male, masculine, mechanical, mighty, militant, military, muscular, neanderthal, notorious, oblivious, offensive, official, outspoken, patriotic, physical, political, powerful, pragmatic, primitive, principal, pro, prominent, prosperous, proud, racist, rational, rebellious, reckless, relentless, republican, rough, ruthless, sarcastic, scary, scientific, selfish, senior, severe, solid, solitary, solo, spartan, stern, stocky, stoic, stout, straightforward, strong, suspect, tall, technical, top, tough, tribal, unafraid, unbeaten, unclean, unconcerned, uncooperative, undefeated, undercover, uninvolved, unrepentant, unstoppable, unsympathetic

Adjectives, Masculine ($2.01 < t < 3.50, 0.001 < p < 0.05$)

aboriginal, addicted, adept, affluent, aged, agnostic, angry, antagonistic, aristocratic, autistic, bad, bankrupt, berserk, biased, brilliant, calm, careless, certain, civilian, cold, communist, confident, confrontational, controversial, cool, culpable, cunning, cynical, despondent, determined, diabetic, dismissive, disruptive, distant, distinguished, dull, employed, empty, evil, explicit, fast, foremost, funny, furious, gay, generic, grumpy, head, heterosexual, homeless, homosexual, honorable, hungry, illegal, illiterate, incumbent, infamous, injured, insistent, irate, irresponsible, lame, lazy, legendary, liable, logical, lone, mad, main, malicious, marked, mean, mischievous, negligent, noble, objective, opposed, orthodox, outside, persistent, pessimistic, pioneering, practical, premier, privileged, productive, professional, punishable, qualified, quick, radical, reasonable, regular, renowned, representative, resolute, respected, retired, rich, robust, rude, rural, sadistic, sane, senile,

serious, sharp, simple, sinful, singular, stable, standard, steadfast, steady, straight, strict, stubborn, successful, suited, sure, tense, terrible, thirsty, traditional, troublesome, ugly, unaffected, unattached, uncivil, unfair, unfaithful, uniformed, unreliable, unscrupulous, unsuitable, unworthy, urban

Adjectives, Slightly Masculine ($1.30 < t < 2.01, 0.05 < p < 0.2$)

accurate, active, agile, alert, alone, averse, awful, awkward, bilingual, bored, broke, collegiate, comatose, common, credible, crucial, deceased, deficient, diligent, dishonest, disposed, distracted, dynamic, eligible, energetic, flawed, guilty, immoral, important, indigenous, indispensable, inept, informed, insolvent, instrumental, intent, international, jobless, junior, knowledgeable, migrant, native, obnoxious, obscure, ordinary, overseas, partisan, pissed, prestigious, ready, realistic, reclusive, recognizable, reputable, resourceful, retarded, revolutionary, ridiculous, right, scholarly, secure, sexual, skilled, stereotypical, sterile, sufficient, superior, unattractive, unaware, unfit, unfortunate, unlucky, unpleasant, unprepared, unresponsive, untrustworthy

Adjectives, Neutral ($-1.30 < t < 1.30, p > 0.2$)

academic, acceptable, accomplished, accountable, accustomed, admirable, advanced, affected, aggrieved, agitated, alien, alive, allergic, amateur, ambivalent, annoyed, annoying, approachable, apt, argumentative, articulate, average, awake, awesome, bedridden, blind, blue, boring, bossy, bright, broken, busy, candid, capable, captive, carefree, central, charming, childish, clear, clever, clumsy, comfortable, competent, complicit, confused, connected, consistent, conspicuous, contemptuous, conventional, convinced, correct, cowardly, crippled, critical, curious, dead, deaf, decent, defensive, defunct, delusional, democratic, deranged, desperate, destitute, devious, different, difficult, dignified, disabled, disappointed, disenchanted, disgusted, disillusioned, displeased, dissatisfied, distinct, distrustful, disturbed, doubtful, dubious, dumb, dyslexic, eager, early, earnest, easy, eccentric, educated, effective, efficient, elderly, eminent, enigmatic, entertaining, entitled, essential, estranged, ethnic, evasive, excellent, exceptional, exclusive, exempt, exhausted, experienced, extraordinary, faceless, fair, fallen, false, famous, fat, favourable, fictional, fierce, fit, flat, fluent, foolish, foreign, forthcoming, fortunate, free, frustrated, full, fun, gifted, grand, great, handicapped, healthy, holy, hypothetical, iconic, ideal, idealistic, identifiable, idle, illegitimate, illustrious, immortal, impaired, impartial, impatient, impossible, impoverished, impressive, impulsive, inactive, incompetent, inconsistent, incorrect, indebted, indignant, individual, ineffective, inexperienced, infected, infirm, insane, inspirational, inspiring, insured, integral, intellectual, intelligent, interesting, invisible, irrelevant, irritated, isolated, jealous, just, keen, key, legitimate, limited, literate, lost, low, loyal, lucky, magnificent, masked, medical, miserable, missing, mistaken, mobile, moderate, mortal, multilingual, muslim, nasty, nearby, negative, neutral, normal, notable, noteworthy, nuts, obese, odd, ok, old, original, outgoing, outstanding, overprotective, overweight, partial, pathetic, penniless, pious, pivotal, plain, playable,

poor, popular, possessed, possessive, preoccupied, private, privy, problematic, proficient, prolific, promiscuous, promising, protective, prudent, psychotic, regent, relaxed, relevant, reliable, reliant, remarkable, repentant, resentful, resistant, respectable, respectful, responsible, restless, reticent, righteous, royal, satisfied, savvy, schizophrenic, secular, shallow, shrewd, sick, sickly, significant, silent, single, skeptical, smart, sober, socialist, sound, specific, stateless, strange, stressed, stupid, suicidal, suitable, suspicious, talented, tendentious, terrific, thorough, transsexual, troubled, truthful, unavailable, unbiased, uncertain, uncomfortable, unconscious, unconvinced, undrafted, uneasy, uneducated, unemployed, unhappy, unharmed, unhurt, unimportant, unimpressed, uninjured, uninsured, uninterested, unmarried, unmasked, unnotable, unpopular, unqualified, unreasonable, unremarkable, unsuccessful, untouchable, untrained, unusual, unwanted, unwelcome, upbeat, upright, useful, useless, vague, vain

Adjectives, Slightly Feminine ($-2.01 < t < -1.30, 0.05 < p < 0.2$)

accessible, amazing, apprehensive, astute, bitter, celebrated, civil, conscious, content, cooperative, dedicated, elusive, envious, exciting, faithful, fake, familiar, fascinating, favourite, fictitious, formal, good, ill, incredible, inferior, insignificant, late, likeable, minor, mute, naked, necessary, paranoid, rare, reluctant, responsive, safe, samaritan, selective, sensible, silly, startled, studious, tired, unbalanced, underage, unnoticed, unpredictable, unwell

Adjectives, Feminine ($-3.50 < t < -2.01, 0.001 < p < 0.05$)

abused, agreeable, amenable, answerable, appreciative, appropriate, attractive, barren, benevolent, bisexual, careful, clueless, committed, compatible, complicated, complimentary, conscientious, contemporary, conversant, crazy, deep, delighted, delirious, depressed, deserving, detailed, devoted, disadvantaged, distraught, distressed, diverse, dizzy, eloquent, fantastic, flamboyant, flexible, fresh, genuine, glad, grammatical, grateful, green, gullible, happy, hesitant, homesick, honest, horrified, imaginary, incapable, indecisive, ineligible, infertile, lean, lenient, liberated, lonely, loose, manipulative, married, mature, memorable, meticulous, misunderstood, modern, moral, musical, mysterious, natural, naughty, neurotic, nostalgic, obedient, observant, obsessive, offended, organized, overjoyed, pale, patient, perceptive, pleasant, pleased, positive, progressive, psychic, quiet, religious, reminiscent, remorseful, reserved, round, sacred, sad, saintly, scared, secretive, sincere, slight, slow, sociable, sorry, special, spirited, stunned, superstitious, susceptible, trusted, trustworthy, unclear, underprivileged, unique, unpaid, unsatisfied, unstable, unsure, unsuspecting, unwilling

Adjectives, Strongly Feminine ($t < -3.50, p < 0.001$)

affectionate, afraid, amiable, anxious, apologetic, artistic, ashamed, attached, attentive, beautiful, beloved, blonde, caring, cautious, charitable, cheerful, clean, compassion-

ate, complex, concerned, considerate, courteous, creative, cultured, cute, dear, delicate, dependent, desirable, desirous, devout, divine, domestic, dressed, ecstatic, elegant, embarrassed, emotional, engaged, enthusiastic, excited, fascinated, fashionable, fearful, feeble, feisty, female, feminine, feminist, fertile, fine, flirtatious, fond, fragile, frail, friendly, frightened, generous, gentle, glamorous, gorgeous, graceful, gracious, haughty, hearted, helpful, helpless, hopeful, hot, humble, hysterical, imaginative, inclusive, innocent, insecure, intimate, kind, lesbian, liberal, light, lightweight, little, lively, lovable, lovely, loving, meek, merciful, mild, mindful, modest, naive, needy, nervous, nice, nude, open, optimistic, passionate, passive, peaceful, perfect, petite, playful, polite, powerless, pregnant, pretty, proper, provocative, pure, receptive, romantic, secret, seductive, selfless, sensitive, sexy, short, shy, skinny, slender, slim, small, social, soft, sophisticated, spiritual, spoiled, stylish, submissive, subordinate, supportive, surprised, sweet, sympathetic, thankful, thin, thoughtful, timid, tiny, tolerant, topless, unarmed, undecided, upset

Verbs, Strongly Masculine ($t > 3.50, p < 0.001$)

abandon, abuse, advance, aim, amass, anchor, appoint, apprentice, arm, arrest, assault, assemble, assert, attack, avenge, bat, battle, beat, bid, block, break, build, build up, burn, bury, campaign, captain, carve, catch, challenge, champion, charge, chase, clash, climb, coach, collide, command, commission, compete, confront, conquer, construct, contract, control, convict, crash, damage, declare, defeat, defend, deliver, demand, deploy, destroy, devise, dig, direct, discharge, discover, dispatch, dispute, dominate, draft, drag, drink, drive, drop out, earn, elect, eliminate, embark, enforce, enlist, erect, establish, excel, execute, exile, experiment, explore, father, fight, finance, fire, fix, force, forge, fund, govern, guard, harass, head, hire, hit, hunt, impose, imprison, inaugurate, incorporate, increase, injure, insert, institute, intercept, invade, invent, invest, investigate, jump, kick, kidnap, kill, knight, knock out, launch, lead, lift, load, loan, log, major, manage, master, mount, murder, occupy, officiate, operate, oppose, ordain, order, oversee, overtake, own, patent, patrol, pay, pilot, pioneer, pitch, preach, preside, proclaim, promote, propose, prosecute, protect, provoke, pull, pull out, punch, punish, pursue, push, race, rank, rap, rebuild, recruit, reign, repair, restrict, retire, risk, rob, row, rule, sack, sail, score, scrap, secure, seize, sell, shoot, shoot down, smash, smoke, spearhead, stab, stalk, steal, steer, stick, strengthen, strike, strike out, sue, supervise, survive, suspend, swear, tackle, take, take on, take over, terminate, threaten, throw, tie, title, top, toss, tow, track, train, transport, umpire, vandalize, venture, violate, wield

Verbs, Masculine ($2.01 < t < 3.50, 0.001 < p < 0.05$)

accomplish, account, accumulate, acquire, announce, ascend, assess, assign, attain, attempt, ban, base, bear, bestow, cap, capture, carry, carry out, cause, chair, cheat, commute, condemn, conduct, consume, contest, crush, cut, cut off, deal, debate, decide, decommission, demonstrate, deny, depart, die, disrupt, distance, distinguish, divide, do, don, drop, dump, eat, employ, endow, estimate, exercise, expand, expel, film, fly, focus,

frame, gain, generate, get, go, grab, grant, grow up, guide, hang, implement, induct, influence, inspect, insult, issue, journey, judge, land, lease, lie, lobby, locate, lock, march, mark, mentor, modify, move, move up, negotiate, nickname, nominate, originate, overcome, owe, pick up, poll, practice, press, proceed, prove, provide, put in, regain, remove, rescue, research, resign, resolve, resume, retrieve, review, ride, rise, ruin, rush, seal, sentence, set off, shift, shout, sign, sign up, smuggle, sneak, speak out, specialize, spread out, stand, stand up, start, station, step, succeed, summon, surpass, take up, target, team, test, total, track down, trade, trick, turn, turn up, undertake, upload, utilize, waste

Verbs, Slightly Masculine ($1.30 < t < 2.01, 0.05 < p < 0.2$)

access, achieve, address, answer, award, become, begin, blackmail, board, cease, check, claim, consecrate, consolidate, contribute, cross, delete, designate, dismiss, distribute, draw up, dub, encounter, engrave, escape, evade, exit, expose, extend, function, go up, hail, hand over, handle, indicate, inherit, initiate, intend, interview, justify, knock, leave, measure, migrate, narrate, obtain, outline, peak, permit, pledge, print, prompt, put down, rally, recommend, register, release, renounce, represent, require, rid, separate, set out, sponsor, start out, step down, stop, supply, suppress, take out, testify, tour, transfer, travel, turn in, turn out, undergo, undo, warn

Verbs, Neutral ($-1.30 < t < 1.30, p > 0.2$)

accord, accuse, act, adapt, add, administer, admire, advertise, advise, afford, agree, align, allow, ally, alter, alternate, apply, approach, approve, argue, associate, assume, attribute, author, average, avoid, back, back up, bequeath, betray, bind, blame, blank, blow, borrow, bowl, break off, break up, bring, bring in, bring out, broadcast, call up, cancel, cast, characterize, choose, cite, classify, clear, close, collaborate, collapse, collect, combine, come, come in, come on, come out, come up, commence, comment, commit, compile, complete, compose, conceal, concentrate, conclude, confer, confine, confirm, consult, contact, contain, contrast, convert, convey, convince, coordinate, copy, correct, count, cover, create, credit, criticize, crown, cruise, decline, deem, defect, define, denounce, depict, derive, descend, describe, deserve, determine, develop, diagnose, disagree, disappear, disclose, disguise, dislike, display, dispose, divorce, document, double, draw, drown, edit, elevate, emerge, emigrate, emphasize, end, end up, endorse, engage, enroll, enter, entrust, envision, evacuate, examine, exhibit, experience, explain, face, facilitate, fail, favor, feature, fill, fill in, find, finish, fit, follow up, forget, form, formulate, found, free, fulfil, further, garner, gather, go on, go out, graduate, grow, hand, hear, hold, hop, hurt, identify, ignore, illustrate, immerse, immigrate, impress, improve, inform, inject, insist, instruct, integrate, interpret, interrupt, introduce, join, keep, keep up, know, lack, lay out, lecture, lend, limit, link, live, look, lose, lose out, lower, lure, maintain, make, manipulate, match, matriculate, mean, merge, miss out, mistake, move on, name, note, notify, offer, opt, paint, part, participate, pass, pass on, pen, perform, persuade, pick, place, plan, play, point, point out, possess, post, predict, present, preserve, produce,

progress, pronounce, protest, publish, put, qualify, question, quit, quote, raise, reach, react, read, record, recount, recover, refer, reference, reform, refuse, regard, relocate, rely, remark, remarry, render, renew, rent, repeat, replace, report, reprise, resemble, reside, resist, resort, respect, rest, restore, retain, return, revert, revise, revive, reward, run, screen, search, see, seek, seek out, select, send, send out, set, set up, shake, show, show up, sink, skate, skip, sleep, solve, sound, speak, split, spot, stand down, state, stay on, steam, stress, struggle, study, stumble, subject, suggest, surprise, survey, suspect, sustain, switch, tag, take off, tally, team up, tell, term, think, transform, transition, translate, trap, try, try out, turn down, turn over, uncover, update, urge, use, vote, wake up, wander, weigh

Verbs, Slightly Feminine ($-2.01 < t < -1.30, 0.05 < p < 0.2$)

acknowledge, advocate, appear, arrive, attend, cede, change, compare, connect, continue, cultivate, debut, enable, endure, enhance, enjoy, ensure, entertain, equal, exchange, fall, find out, flee, give up, guest, host, involve, lay, let, meet, mix, move in, object, observe, perceive, portray, pour, prepare, prevent, put up, reappear, recall, recite, recognize, recreate, reduce, regret, reject, relate, retreat, reveal, sacrifice, satisfy, save, say, sit, spell, spread, star, tear, trace, visit, wait, want, watch

Verbs, Feminine ($-3.50 < t < -2.01, 0.001 < p < 0.05$)

absorb, accept, accompany, admit, age, attach, audition, befriend, behave, believe, belong, benefit, book, bring up, buy, call, complain, concede, concern, confess, confuse, consider, correspond, date, dedicate, devote, donate, educate, encourage, escort, expect, fake, file, help out, hide, highlight, imagine, include, inspire, interact, intern, label, learn, liken, list, meet up, mention, need, notice, organize, partner, perfect, petition, poison, prefer, promise, purchase, put on, realize, receive, reconcile, rejoin, remain, remind, rename, reply, request, respond, reunite, serve, shape, side, sit out, slap, slip, stay, succumb, suffer, supplement, support, take in, talk, tend, touch, trust, unite, upset, walk, wear

Verbs, Strongly Feminine ($t < -3.50, p < 0.001$)

adopt, aid, apologize, appeal, appreciate, arrange, ask, assist, assure, attract, await, beg, care, celebrate, communicate, conceive, confide, cooperate, cry, dance, decorate, depend, design, discuss, dream, dress, embrace, express, fear, feed, feel, follow, forgive, give, greet, help, invite, kiss, like, listen, love, make up, marry, miss, model, open, open up, photograph, plant, plead, pose, praise, pray, reflect, relieve, remember, reorganize, seduce, settle, share, sing, spend, strip, submit, substitute, surrender, teach, thank, treat, understand, view, voice, volunteer, wash, welcome

A.5 Survey for Gender Stereotype in Natural Language

Instruction

The goal of this study is to understand **how people understand gender stereotypes**. Your task is to find articles on the Internet **that describe people who exemplify or contradict common gender stereotypes of US society**. Here, *gender stereotypes* are general impressions that society places on men and women, **not necessarily stereotypes that you hold yourself** for men and women.

For example, if you think men are stereotypically viewed in society as powerful and rich, and you are asked to find an article about stereotypical men, you might want to look for articles that describe a male Wall Street trader.

The article can come from news webpages, wikipedia pages, online social media, obituaries, self-description in public resumes, etc. You will need to choose 4 different articles, and we will pay you $0.75 for each article. You will be given instructions for each article that we want you to find. Please read these instructions carefully each time because they may vary for each article.

Questions

Please search the web and choose an article so that:

1) The article contains description of a MAN *(or WOMAN)*. It could be related to his *(or her)* appearance, characteristics, habit, or a story about the person.

2) His *(or Her)* characteristics and behaviors are CONSISTENT with *(or CONTRADICT)* a typical man *(or woman)* according to stereotypes in society.

- Please paste the paragraphs that are related to the person (if only a few paragraphs are related to the person, please ONLY copy and paste these paragraphs).
- Please paste the link to the webpage where you find the article.
- Please briefly state why the person in the paragraphs contradict common gender stereotypes.

Each user is assigned 4 articles, with different combinations of requirements.

Demographics

1. How old are you?

- 18 - 19
- 20 - 30
- 31 - 40
- 41 - 50
- 51 - 60
- 60+

2. What is your gender?

- Female
- Male
- Other

3. What is your ethnicity?

- Black or African American
- Asian
- Hispanic or Latino
- Pacific Islander
- White
- Other

4. What is the highest level of education you completed?

- Less than a high school diploma
- High school or equivalent (GED)
- Some college (no degree)
- Associate's degree (e.g., AA, AS)
- Bachelor's degree (e.g., BA, BS)
- Master's degree (e.g. MA, MS)
- Advanced graduate work or Ph.D. (e.g., PhD, MD, DDS)

5. For quality control, please choose A and D for this question (please choose both):

- A
- B
- C
- D
- E

6. Which of the following best describes your current occupation?

- Student
- Full-time employment
- Part-time employment
- Self-employed
- Unemployed, looking for work
- Unemployed, not looking for work
- Retired

Bibliography

[1] G. A. Miller, *The magical number seven, plus or minus two: Some limits on our capacity for processing information.*, *Psychological review* **63** (1956), no. 2 81.

[2] P. C. Wason, *On the failure to eliminate hypotheses in a conceptual task*, *Quarterly journal of experimental psychology* **12** (1960), no. 3 129–140.

[3] M. Foschi, L. Lai, and K. Sigerson, *Gender and double standards in the assessment of job applicants*, *Social Psychology Quarterly* **57** (1994), no. 4 326–339.

[4] A. Tversky and D. Kahneman, *Judgment under uncertainty: Heuristics and biases*, *science* **185** (1974), no. 4157 1124–1131.

[5] M. G. Haselton, D. Nettle, and D. R. Murray, *The evolution of cognitive bias*, *The handbook of evolutionary psychology* (2015) 1–20.

[6] E. Bakshy, S. Messing, and L. A. Adamic, *Exposure to ideologically diverse news and opinion on facebook*, *Science* **7** (May, 2015) 1130–1132.

[7] S. Yardi and D. Boyd, *Dynamic debates: An analysis of group polarization over time on twitter*, *Bulletin of Science, Technology, and Society* **30** (2010), no. 5 316–327.

[8] C. Sustein, *Republic.com*. Princeton University Press, 2001.

[9] R. Prechter, *The Wave Principle of Human Social Behavior*, p. 152153. New Classics Library, 1999.

[10] A. Brown and E. Patten, "The narrowing, but persistent, gender gap in pay." `http://www.pewresearch.org/fact-tank/2017/04/03/gender-pay-gap-facts/`, 2017.

[11] T. W. B. Group, "Many societies gradually moving to dismantle gender discrimination, yet more can be done, says world bank group president jim yong kim." `http://www.worldbank.org/en/news/press-release/2013/09/24/`

societies-dismantle-gender-discrimination-world-bank-group-president-jim-yong-ki
2017.

[12] F. D. Blau and L. M. Kahn, *The gender wage gap: Extent, trends, and explanations*, tech. rep., National Bureau of Economic Research, 2016.

[13] H. Minter, "Why so few women in tech? your answer depends on your gender." https://www.theguardian.com/careers/2016/dec/12/why-so-few-women-in-tech-your-answer-depends-on-your-gender, 2017.

[14] S. D. Solomon, "Why so few women reach the executive rank." https://dealbook.nytimes.com/2013/04/02/why-so-few-women-reach-the-executive-rank/, 2017.

[15] L. S. Gottfredson, *Circumscription and compromise: A developmental theory of occupational aspirations.*, *Journal of Counseling psychology* **28** (1981), no. 6 545.

[16] M. A. Cejka and A. H. Eagly, *Gender-stereotypic images of occupations correspond to the sex segregation of employment*, *Personality and Social Psychology Bulletin* **25** (1999), no. 4 413–423.

[17] D. Neumark, R. J. Bank, and K. D. Van Nort, *Sex discrimination in restaurant hiring: An audit study*, *The Quarterly Journal of Economics* **111** (1996), no. 3 915–941.

[18] D. Gaucher, J. Friesen, and A. C. Kay, *Evidence that gendered wording in job advertisements exists and sustains gender inequality.*, *Journal of Personality and Social Psychology* **101** (2011), no. 1 109.

[19] C. Wagner, D. Garcia, M. Jadidi, and M. Strohmaier, *It's a man's wikipedia? assessing gender inequality in an online encyclopedia*, in *Proc. of ICWSM*, 2015.

[20] J. M. Madera, M. R. Hebl, and R. C. Martin, *Gender and letters of recommendation for academia: agentic and communal differences.*, *Journal of Applied Psychology* **94** (2009), no. 6 1591.

[21] E. Fast, T. Vachovsky, and M. S. Bernstein, *Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community*, in *Proc. of ICWSM*, 2016.

[22] A. Ramakrishna, V. R. Martínez, N. Malandrakis, K. Singla, and S. Narayanan, *Linguistic analysis of differences in portrayal of movie characters*, in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 1669–1678, 2017.

[23] J. M. Twenge, *Changes in masculine and feminine traits over time: A meta-analysis*, *Sex roles* **36** (1997), no. 5-6 305–325.

[24] K. Donnelly and J. M. Twenge, *Masculine and feminine traits on the bem sex-role inventory, 1993–2012: a cross-temporal meta-analysis*, *Sex Roles* (2016) 1–10.

[25] S. Tang, X. Zhang, J. Cryan, M. J. Metzger, H. Zheng, and B. Y. Zhao, *Gender bias in the job market: A longitudinal analysis*, *Proceedings of the ACM on Human-Computer Interaction* **1** (2017), no. CSCW 99.

[26] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, *Deep learning for hate speech detection in tweets*, in *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 759–760, International World Wide Web Conferences Steering Committee, 2017.

[27] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, *Lexicon-based methods for sentiment analysis*, *Computational linguistics* **37** (2011), no. 2 267–307.

[28] T. Davidson, D. Warmsley, M. Macy, and I. Weber, *Automated hate speech detection and the problem of offensive language*, in *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[29] D. Kahneman and A. Tversky, *Subjective probability: A judgment of representativeness*, *Cognitive psychology* **3** (1972), no. 3 430–454.

[30] H. R. Arkes, *Costs and benefits of judgment errors: Implications for debiasing.*, *Psychological bulletin* **110** (1991), no. 3 486.

[31] R. S. Nickerson, *Confirmation bias: A ubiquitous phenomenon in many guises*, *Review of general psychology* **2** (1998), no. 2 175–220.

[32] C. G. Lord, L. Ross, and M. R. Lepper, *Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence.*, *Journal of personality and social psychology* **37** (1979), no. 11 2098.

[33] M. Snyder and N. Cantor, *Testing hypotheses about other people: The use of historical knowledge*, *Journal of Experimental Social Psychology* **15** (1979), no. 4 330–342.

[34] E. Gilbert, T. Bergstrom, and K. Karahalios, *Blogs are echo chambers: Blogs are echo chambers*, in *Proc. of HICCS*, 2009.

[35] C. Grevet, L. G. Terveen, and E. Gilbert, *Managing political differences in social media*, in *Proc. of CSCW*, 2014.

[36] R. K. Garrett, *Echo chambers online?: Politically motivated selective exposure among internet news users*, *Journal of Computer-Mediated Communication* **14** (2009), no. 2 265–285.

[37] M. M. Pompian, *Behavioral finance and wealth management: how to build investment strategies that account for investor biases*, vol. 667. John Wiley & Sons, 2011.

[38] A. Hergovich, R. Schott, and C. Burger, *Biased evaluation of abstracts depending on topic and conclusion: Further evidence of a confirmation bias within scientific psychology*, *Current Psychology* **29** (2010), no. 3 188–209.

[39] A. M. Colman, *A dictionary of psychology*. Oxford Quick Reference, 2015.

[40] S. S. Sundar, A. Oeldorf-Hirsch, and Q. Xu, *The bandwagon effect of collaborative filtering technology*, in *CHI'08 extended abstracts on Human factors in computing systems*, pp. 3453–3458, ACM, 2008.

[41] S. S. Sundar and C. Nass, *Conceptualizing sources in online news*, *Journal of Communication* **51** (2001), no. 1 52–72.

[42] I. McAllister and D. T. Studlar, *Bandwagon, underdog, or projection? opinion polls and electoral choice in britain, 1979-1987*, *The Journal of Politics* **53** (1991), no. 3 720–741.

[43] R. J. Shiller and J. Pound, *Survey evidence of diffusing of interest among institutional investors*, *Journal of Economic Behavior and Organization* **12** (August, 1989) 47–66.

[44] Y.-F. Chen, *Herd behavior in purchasing books online*, *Computers in Human Behavior* **24** (2008), no. 5 1977–1992.

[45] A. H. Eagly, W. Wood, and A. B. Diekman, *Social role theory of sex differences and similarities: A current appraisal*, *The developmental social psychology of gender* **12** (2000) 174.

[46] A. J. Cuddy, S. T. Fiske, and P. Glick, *When professionals become mothers, warmth doesn't cut the ice*, *Journal of Social issues* **60** (2004), no. 4 701–718.

[47] B. Wittenbrink, C. M. Judd, and B. Park, *Evidence for racial prejudice at the implicit level and its relationship with questionnaire measures.*, *Journal of personality and social psychology* **72** (1997), no. 2 262.

[48] D. M. Mackie, D. L. Hamilton, J. Susskind, and F. Rosselli, *Social psychological foundations of stereotype formation*, *Stereotypes and stereotyping* (1996) 41–78.

[49] D. L. Hamilton and R. K. Gifford, *Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments*, *Journal of Experimental Social Psychology* **12** (1976), no. 4 392–407.

[50] A. G. Greenwald and M. R. Banaji, *Implicit social cognition: attitudes, self-esteem, and stereotypes.*, *Psychological review* **102** (1995), no. 1 4.

[51] D. Kahneman and P. Egan, *Thinking, fast and slow*, vol. 1. Farrar, Straus and Giroux New York, 2011.

[52] S. Sinclair, E. Dunn, and B. Lowery, *The relationship between parental racial attitudes and childrens implicit prejudice*, *Journal of Experimental Social Psychology* **41** (2005), no. 3 283–289.

[53] N. Dasgupta, D. DeSteno, L. A. Williams, and M. Hunsinger, *Fanning the flames of prejudice: the influence of specific incidental emotions on implicit prejudice.*, *Emotion* **9** (2009), no. 4 585.

[54] P. G. Devine, *Stereotypes and prejudice: Their automatic and controlled components.*, *Journal of personality and social psychology* **56** (1989), no. 1 5.

[55] M. R. Banaji and C. D. Hardin, *Automatic stereotyping*, *Psychological science* **7** (1996), no. 3 136–141.

[56] A. G. Greenwald, D. E. McGhee, and J. L. Schwartz, *Measuring individual differences in implicit cognition: the implicit association test.*, *Journal of personality and social psychology* **74** (1998), no. 6 1464.

[57] J. Correll, B. Park, C. M. Judd, B. Wittenbrink, M. S. Sadler, and T. Keesee, *Across the thin blue line: police officers and racial bias in the decision to shoot.*, *Journal of personality and social psychology* **92** (2007), no. 6 1006.

[58] J. L. Eberhardt, P. G. Davies, V. J. Purdie-Vaughns, and S. L. Johnson, *Looking deathworthy: Perceived stereotypicality of black defendants predicts capital-sentencing outcomes*, *Psychological science* **17** (2006), no. 5 383–386.

[59] R. Bloomfield, M. Ohara, and G. Saar, *How noise trading affects markets: An experimental analysis*, *Review of Financial Studies* **22** (January, 2009) 2275–2302.

[60] L. Kogan, S. Ross, J. Wang, and M. Westerfield, *The survival and price impact of irrational traders*, *Journal of Finance* **61** (February, 2006) 195–229.

[61] S. R. Das, *Text and context: Language analytics in finance*, *Foundations and Trends (R) in Finance* **8** (2014), no. 3 145–261.

[62] F. Black, *Noise*, *The journal of finance* **41** (1986), no. 3 529–543.

[63] R. J. Shiller, S. Fischer, and B. M. Friedman, *Stock prices and social dynamics*, *Brookings Papers on Economic Activity* **1984** (1984), no. 2 457–498.

[64] A. S. Kyle, *Continuous auctions and insider trading*, *Econometrica* **53** (November, 1985) 1315–1336.

[65] S. Benartzi and R. H. Thaler, *Nave diversification strategies in defined contribution saving plans*, *American Economic Review* **91** (2001) 7998.

[66] D. Kahneman and A. Tversky, *Prospect theory: An analysis of decision under risk*, *Econometrica* **47** (1979) 263291.

[67] B. G. Malkiel and E. F. Fama, *Efficient capital markets: A review of theory and empirical work*, *The Journal of Finance* **25** (1970), no. 2 383–417.

[68] A. Boutyline and R. Willer, *The social structure of political echo chambers: Variation in ideological homophily in online networks*, *Political Psychology* (May, 2016).

[69] R. Wayman, "Understanding small- and big-cap stocks." Investopedia.com, 2016.

[70] W. Zhang and S. Skiena, *Trading strategies to exploit blog and news sentiment.*, in *Proc. of ICWSM*, 2010.

[71] E. Gilbert and K. Karahalios, *Widespread worry and the stock market*, in *Proc. of ICWSM*, 2010.

[72] J. Bollen, H. Mao, and X. Zeng, *Twitter mood predicts the stock market*, *Journal of Computational Science* **2** (2011), no. 1 1–8.

[73] G. Wang, T. Wang, B. Wang, D. Sambasivan, Z. Zhang, H. Zheng, and B. Y. Zhao, *Crowds on wall street: Extracting value from collaborative investing platforms*, in *Proc. of CSCW*, 2015.

[74] H. Chen, P. De, Y. J. Hu, and B.-H. Hwang, *Wisdom of crowds: The value of stock opinions transmitted through social media*, *Review of Financial Studies* **27** (2014), no. 5 1367–1403.

[75] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, *Comparing and combining sentiment analysis methods*, in *Proc. of COSN*, 2013.

[76] F. Pedregosa *et. al.*, *Scikit-learn: Machine learning in Python*, *Journal of Machine Learning Research* **12** (2011) 2825–2830.

[77] H. He and E. A. Garcia, *Learning from imbalanced data*, in *Proc. of TKDE*, 2009.

[78] R. Lowry, *Concepts and applications of inferential statistics*, .

[79] B. Pang, L. Lee, and S. Vaithyanathan, *Thumbs up?: sentiment classification using machine learning techniques*, in *Proc. of EMNLP*, 2002.

[80] K. Pearson, *Note on regression and inheritance in the case of two parents*, *Proc. of the Royal Society of London* (1895).

[81] O. Arazy, F. Ortega, O. Nov, L. Yeo, and A. Balila, *Functional roles and career paths in wikipedia*, in *Proc. of CSCW*, 2015.

[82] N. Sun, P. P.-L. Rau, and L. Ma, *Understanding lurkers in online communities: A literature review*, *Computers in Human Behavior* **38** (2014) 110–117.

[83] H. Zhu, R. Kraut, and A. Kittur, *Effectiveness of shared leadership in online communities*, in *Proc. of CSCW*, 2012.

[84] R. D. Nolker and L. Zhou, *Social computing and weighting to identify member roles in online communities*, in *Proc. of WI*, 2005.

[85] G. Wang, X. Zhang, S. Tang, H. Zheng, and B. Y. Zhao, *Unsupervised clickstream clustering for user behavior analysis*, in *Proc. of CHI*, May, 2016.

[86] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons, 2009.

[87] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, *Fast unfolding of communities in large networks*, *Journal of statistical mechanics: theory and experiment* **2008** (2008), no. 10.

[88] R. J. Shiller, *Irrational Exuberance*, p. 149153. Princeton University Press, 2000.

[89] R. Demirera and A. M. Kutan, *Does herding behavior exist in chinese stock markets?*, *Journal of International Financial Markets, Institutions and Money* **16** (April, 2006) 123–142.

[90] E. C. Chang, J. W. Cheng, and A. Khorana, *An examination of herd behavior in equity markets: An international perspective*, *Journal of Banking & Finance* **24** (October, 2000) 16511679.

[91] J. D. Piotroski and T. Wong, *Institutions and information environment of chinese listed firms*, in *Capitalizing China*, pp. 201–242. University of Chicago Press, 2012.

[92] S. Shen and B. Goh, "China stock market freezing up as sell-off gathers pace." Reuters, July, 2015.

[93] G. Wildau, "Myth of chinas retail investors understates large players role." Financial Times, July, 2015.

[94] A. J. Davis, *A requiem for the retail investor?*, *VA Law Review* **95** (2009), no. 4 1105–29.

[95] J. J. Murphy, *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. Penguin, 1999.

[96] J. C. Hernandez, "Stock downturn hits chinese investors in the heart, not just the wallet." New York Times, July, 2015.

[97] X. Li and K.-J. Chen, *On the value of technical analysis for stock traders in china*, .

[98] C. J. Burke, P. N. Tobler, W. Schultz, and M. Baddeley, *Striatal bold response reflects the impact of herd information on financial decisions*, *Frontiers of Human Neuroscience* **4** (2010), no. 48. doi:10.3389/fnhum.2010.00048.

[99] R. M. Raafatemail, N. Chater, and C. Frith, *Herding in humans*, *Trends in Cognitive Sciences* **13** (2009), no. 10 420–428. doi:10.1016/j.tics.2009.08.002.

[100] E. Chew, "Does chart analysis really work?." The Wall Street Journal, Nov, 2013.

[101] C. Murphy, "Can technical analysis be called a self-fulfilling prophecy?." http://www.investopedia.com/ask/answers/05/-selffulfillingprophecy.asp, 2014.

[102] F. Garzarelli, M. Cristelli, G. Pompa, A. Zaccaria, and L. Pietronero, *Memory effects in stock price dynamics: evidences of technical trading*, *Scientific reports* **4** (2014).

[103] S. J. Jordan, *Is momentum a self-fulfilling prophecy?*, *Quantitative Finance* **14** (2014), no. 4 737–748.

[104] C.-H. Park and S. H. Irwin, *What do we know about the profitability of technical analysis?*, *Journal of Economic Surveys* **21** (2007), no. 4 786–826.

[105] T. T.-L. Chong and W.-K. Ng, *Technical analysis and the london stock exchange: testing the macd and rsi rules using the ft30*, *Applied Economics Letters* **15** (2008), no. 14 1111–1114.

[106] B. R. Marshall, S. Qian, and M. Young, *Is technical analysis profitable on us stocks with certain size, liquidity or industry characteristics?*, *Applied Financial Economics* **19** (2009), no. 15 1213–1221.

[107] R. Rosillo, D. De la Fuente, and J. Brugos, *Technical analysis and the spanish stock exchange: testing the rsi, macd, momentum and stochastic rules using spanish market companies*, *Applied Economics* **45** (2013), no. 12 1541–1550.

[108] B. R. Marshall, R. H. Cahan, and J. Cahan, *Technical analysis around the world*, *Available at SSRN 1181367* (2010).

[109] J. Hartono and D. Sulistiawan, *The market quality to technical analysis performance: Inter-country analysis*, *Available at SSRN 2530376* (2014).

[110] Q. Ying, D. Kong, and D. Luo, *Investor attention, institutional ownership, and stock return: Empirical evidence from china*, *Emerging Markets Finance and Trade* **51** (2015), no. 3 672–685.

[111] T. T.-L. Chong, T.-H. Lam, and I. K.-M. Yan, *Is the chinese stock market really inefficient?*, *China Economic Review* **23** (2012), no. 1 122–137.

[112] M. Hanrahan, "China stock market: Can reforms prevent more stock chaos?." International Business Times, January, 2016. http://www.ibtimes.com/china-stock-market-can-reforms-prevent-more-stock-chaos-2258363.

[113] M. Sadker and D. Sadker, *Failing at fairness: How America's schools cheat girls.* Simon and Schuster, 2010.

[114] L. A. McLoughlin, *Spotlighting: Emergent gender bias in undergraduate engineering education*, *Journal of Engineering Education* **94** (2005), no. 4 373.

[115] J. Clark Blickenstaff*, *Women and science careers: leaky pipeline or gender filter?*, *Gender and education* **17** (2005), no. 4 369–386.

[116] K. Holtzblatt, A. Balakrishnan, T. Effner, E. Rhodes, and T. Tuan, *Beyond the pipeline: addressing diversity in high tech*, in *Proc. of CHI Extended Abstract*, 2016.

[117] D. Tomaskovic-Devey, *Gender & racial inequality at work: The sources and consequences of job segregation.* No. 27. Cornell University Press, 1993.

[118] K. L. Cropsey, S. W. Masho, R. Shiang, V. Sikka, S. G. Kornstein, and C. L. Hampton, *Why do faculty leave? reasons for attrition of women and minority faculty from a medical school: four-year results*, *Journal of Women's Health* **17** (2008), no. 7 1111–1118.

[119] C. M. Marlowe, S. L. Schneider, and C. E. Nelson, *Gender and attractiveness biases in hiring decisions: Are more experienced managers less biased?*, *Journal of Applied Psychology* **81** (1996), no. 1 11.

[120] R. E. Steinpreis, K. A. Anders, and D. Ritzke, *The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study*, *Sex Roles* **41** (1999), no. 7 509–528.

[121] C. L. Ridgeway, *Interaction and the conservation of gender inequality: Considering employment*, *American Sociological Review* **62** (1997), no. 2 218–235.

[122] M. Bertrand and S. Mullainathan, *Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination*, *The American Economic Review* **94** (2004), no. 4 991–1013.

[123] P. A. Riach and J. Rich, *An experimental investigation of sexual discrimination in hiring in the english labor market*, *The B.E. Journal of Economic Analysis & Policy* **5** (2006), no. 2 1–22.

[124] I. Askehave and K. K. Zethsen, *Gendered constructions of leadership in danish job advertisements*, *Gender, Work & Organization* **21** (2014), no. 6 531–545.

[125] S. L. Bem and D. J. Bem, *Does sex-biased job advertising "aid and abet" sex discrimination?*, *Journal of Applied Social Psychology* **3** (1973), no. 1 6–18.

[126] A. Hannák, C. Wagner, D. Garcia, A. Mislove, M. Strohmaier, and C. Wilson, *Bias in online freelance marketplaces: evidence from taskrabbit and fiverr*, in *Proc. of CSCW*, 2017.

[127] U. B. of Labor Statistics, "Median weekly earnings of full-time wage and salary workers by detailed occupation and sex." `https://www.bls.gov/cps/aa1998/CPSAAT39.PDF`, 1998.

[128] U. B. of Labor Statistics, "Median weekly earnings of full-time wage and salary workers by detailed occupation and sex." `https://www.bls.gov/cps/cpsaat39.htm`, 2017.

[129] P. E. Gabriel and S. Schmitz, *Gender differences in occupational distributions among workers*, *Monthly Labor Review* **130** (2007) 19.

[130] L. E. Atwater, J. F. Brett, D. Waldman, L. DiMare, and M. Hayden, *Men's and women's perceptions of the gender typing of management subroles*, *Sex Roles* **50** (2004), no. 3 191–199.

[131] K. D. Joshi and N. L. Schmidt, *Is the information systems profession gendered?: characterization of is professionals and is career*, *ACM SIGMIS Database* **37** (2006), no. 4 26–41.

[132] E. M. Trauth, C. C. Cain, K. Joshi, L. Kvasny, and K. M. Booth, *The influence of gender-ethnic intersectionality on gender stereotypes about it skills and knowledge*, *ACM SIGMIS Database* **47** (2016), no. 3 9–39.

[133] S. J. Ceci, D. K. Ginther, S. Kahn, and W. M. Williams, *Women in academic science a changing landscape*, *Psychological Science in the Public Interest* **15** (2014), no. 3 75–141.

[134] D. I. Miller and J. Wai, *The bachelor's to ph. d. stem pipeline no longer leaks more women than men: a 30-year analysis*, *Frontiers in psychology* **6** (2015) 37.

[135] S. F. Way, D. B. Larremore, and A. Clauset, *Gender, productivity, and prestige in computer science faculty hiring networks*, in *Proc. of WWW*, 2016.

[136] M. L. Newman, C. J. Groom, L. D. Handelman, and J. W. Pennebaker, *Gender differences in language use: An analysis of 14,000 text samples*, *Discourse Processes* **45** (2008), no. 3 211–236.

[137] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, *Man is to computer programmer as woman is to homemaker? debiasing word embeddings*, in *Advances in Neural Information Processing Systems*, pp. 4349–4357, 2016.

[138] T. Schmader, J. Whitehead, and V. H. Wysocki, *A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants*, *Sex Roles* **57** (2007), no. 7-8 509–514.

[139] S. Sczesny, M. Formanowicz, and F. Moser, *Can gender-fair language reduce gender stereotyping and discrimination?*, *Frontiers in Psychology* **7** (2016) 25.

[140] L. K. Horvath and S. Sczesny, *Reducing women's lack of fit with leadership positions? effects of the wording of job advertisements*, *European Journal of Work and Organizational Psychology* **25** (2016), no. 2 316–328.

[141] C. C. Aggarwal and C. Zhai, *A survey of text classification algorithms*, *Mining text data* (2012) 163–222.

[142] S. L. Bem, *The measurement of psychological androgyny*, *Journal of Consulting and Clinical Psychology* **42** (1974), no. 2 155–162.

[143] P. Rosenkrantz, S. Vogel, H. Bee, I. Broverman, and D. M. Broverman, *Sex-role stereotypes and self-concepts in college students.*, *Journal of Consulting and Clinical Psychology* **32** (1968), no. 3 287.

[144] J. T. Spence, R. Helmreich, and J. Stapp, *Ratings of self and peers on sex role attributes and their relation to self-esteem and conceptions of masculinity and femininity.*, *Journal of Personality and Social Psychology* **32** (1975), no. 1 29.

[145] S. A. Schullo and B. L. Alperson, *Interpersonal phenomenology as a function of sexual orientation, sex, sentiment, and trait categories in long-term dyadic relationships.*, *Journal of Personality and Social Psychology* **47** (1984), no. 5 983.

[146] P. Glick and S. T. Fiske, *The ambivalent sexism inventory: differentiating hostile and benevolent sexism*, *Journal of Personality and Social Psychology* **70** (1996), no. 3 491–512.

[147] R. Johnson and T. Zhang, *Supervised and semi-supervised text categorization using lstm for region embeddings*, in *International Conference on Machine Learning*, pp. 526–534, 2016.

[148] M. Menegatti and M. Rubini, *Gender bias and sexism in language*, 2017.

[149] A. Maass and L. Arcuri, *Language and stereotyping*, *Stereotypes and stereotyping* (1996) 193–226.

[150] J. T. Spence, R. L. Helmreich, and J. Stapp, *The personal attributes questionnaire: A measure of sex role stereotypes and masculinity-femininity*, *Journal Supplement Abstract Service Catalog of Selected Documents in Psychology* **4** (1974).

[151] M. L. Dean and C. C. Tate, *Extending the legacy of sandra bem: Psychological androgyny as a touchstone conceptual advance for the study of gender in psychological science*, *Sex Roles* **76** (2017), no. 11-12 643–654.

[152] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, *Gender bias in coreference resolution: Evaluation and debiasing methods*, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 2, 2018.

[153] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach, *Women also snowboard: Overcoming bias in captioning models*, in *European Conference on Computer Vision*, pp. 793–811, 2018.

[154] J. H. Park, J. Shin, and P. Fung, *Reducing gender bias in abusive language detection*, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2799–2804, 2018.

[155] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang, *Learning gender-neutral word embeddings*, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4847–4853, 2018.

[156] Student, *The probable error of a mean*, *Biometrika* (1908) 1–25.

[157] S. Mohammad, *Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words*, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 174–184, 2018.

[158] A. Caliskan, J. J. Bryson, and A. Narayanan, *Semantics derived automatically from language corpora contain human-like biases*, *Science* **356** (2017), no. 6334 183–186.

[159] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, *Distributed representations of words and phrases and their compositionality*, in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[160] J. Pennington, R. Socher, and C. Manning, *Glove: Global vectors for word representation*, in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[161] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, *Enriching word vectors with subword information*, *Transactions of the Association for Computational Linguistics* **5** (2017) 135–146.

[162] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[163] Y. Yang and J. O. Pedersen, *A comparative study on feature selection in text categorization*, in *ICML*, vol. 97, p. 35, 1997.

[164] D. Stahlberg, F. Braun, L. Irmen, and S. Sczesny, *Representation of the sexes in language*, *Social communication* (2007) 163–187.

[165] M. C. Hamilton, *Using masculine generics: Does generic he increase male bias in the user's imagery?*, *Sex roles* **19** (1988), no. 11-12 785–799.

[166] F. Pratto, J. D. Korchmaros, and P. Hegarty, *When race and gender go without saying*, *Social Cognition* **25** (2007), no. 2 221–247.