NORTHWESTERN UNIVERSITY

Harnessing Web Information Sources to Predict Events

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Computer Science

By

Mohammed A. Alam

EVANSTON, ILLINOIS

September 2020

Abstract

Harnessing Web Information Sources to Predict Events

Mohammed A. Alam

Search engines and social media are two ubiquitous modes of accessing Web information, and they dictate what information people view, influencing their thoughts and beliefs and potentially shaping their opinions about news and facts. Network effects propagate the beliefs, often magnified, disseminating information rapidly and leaving little time for fact checks. This gives rise to a problem: it is not only information that spreads, but also misinformation. The resulting far-reaching impacts include misinformed prognostications that can lead to further ill effects.

Considerable attention has been given to the “cleanup” of the Web, focusing on the common purpose of providing accountability to statements made online. However, the size and the growth of the Web make it challenging to characterize Web information or to separate facts from lies, resulting in people’s thoughts and actions that can be void of truth.

In this dissertation, we address the problem by using methods based on our thesis that Web information sources can be harnessed to synthesize accurate predictions of events in an attempt to arrive at the truth. Instead of validating every piece of information for provenance, which can be recursive and quickly become intractable, we adopt an approach

that embraces noise in information and relies on the wisdom of crowds to derive accurate predictions from data.

Toward our goal, we first characterize a particular bias of Web search engine results: the degree to which differences across engines' rankings correlate with features of the ranked content, including point of view and advertisements. We develop *PAWS*—Platform for Analyzing Web Search engines—to study Google and Bing, and we find no evidence that the engines emphasize results expressing positive orientation toward the engine company's products. We do find that they emphasize particular news sites and that they also favor pages containing their company's advertisements, as opposed to competitors'.

Next, we use sports predictions from Twitter crowds to study methods for predicting game outcomes. We show that the wisdom of crowds and machine learning can lead to accurate predictions for certain games, and that features pertaining to the crowds can be leveraged for the purpose of prediction. We test similar approaches using Earnings Per Share and Revenue predictions from financial prediction platform, Estimize, and show that our methods have potential applicability across domains for deriving the truth through predictions.

Acknowledgments

I would like to start by acknowledging my phenomenal adviser and Chair of my Thesis Committee, Dr. Douglas C. Downey. Doug has been an absolutely fantastic adviser to me during my doctoral pursuit, and I cannot thank him enough. I owe a great deal of my learning—in Artificial Intelligence and in Computer Science in general—to Doug, whose whiteboard snapshots will continue to serve me as reference for years to come. Thank you, Doug, for always taking the time to distill difficult concepts for me. The patience with which you frequently took questions from me—often the same question over and over (and over) again until the concept was imprinted in my brain—is simply remarkable! Thank you for readily accepting me as a researcher in your group, for believing in me since, for your constant support and encouragement, and for guiding me throughout my entire doctoral journey. Thank you for everything you have done to help me to get here! And I wish you the very best in your continued amazing research in AI.

I would like to thank the rest of my Committee, Dr. Lawrence A. Birnbaum and Dr. Bryan A. Pardo. Larry's comparison of engineering and science in one class was, for me as an engineer who had just started graduate school to become a scientist, foundational for a lot of my thoughts about AI. Larry's insightful questions on many occasions taught me how to think from different angles. From sitting in Bryan's Introduction to AI class on my first day at Northwestern through a total of four of his AI classes, I not only learned AI from Bryan, but his way of teaching inspired how I would love to teach in my future academic pursuits. Bryan has also always been extremely supportive and has endured my many questions, taking the time to make concepts digestible for me. Thank you, Larry and Bryan, for always being there, and for vetting my work as members of my Committee!

Now, I would like to acknowledge some very special people: my loving wife, Nuzhat Maherin, and my amazing daughters, Liyana and Rosie, without whom no part of my doctoral pursuit would be even imaginable. I have managed to write a few words toward my dissertation, but words fail me in describing the tremendous amount of sacrifice they have made for me during these last few years of my graduate school. Two of them were born during this time, and they have shown a capacity for patience, understanding, and caring that is well beyond their age. I am simply the luckiest man around to have received the kind of relentless support that my lovely wife Nuzhat has given me while she has worked hard to make sure I could focus on studying. My sweet girls have postponed many a play time or walk or bike ride to make sure I could finish my research or my dissertation. Having spent years trying to use machine learning to predict things, I still find it baffling how my wife and my girls could anticipate my need for caffeine and bring me a cup of tea or coffee with a snack exactly at the right moment. However, the words of comfort they have offered during my occasional bouts of despair have provided more energy and motivation than any amount of caffeine could do. To my wife and children, I want to say, thank you so very much for the relentless support and love you have given me and for everything you have done for me so I could earn my PhD!

I would like to thank my dear mother, Anwara Begum, and my dear father, Mohammad Abdus Samad, for being my parents and for being there with all the love in the world from the moment I was born to the moment my father cried on the phone when he heard I had passed my Thesis Defense, and they continue to be there for me. What my mother means to me and what she has done are captured in that single unigram: *mother.* No words will do justice to what she has done for me, how much she has motivated me, and how much love she has given me. My father, a brilliant Civil Engineer, was my inspiration for becoming an engineer. Due to the need to take care of family, he was unable to pursue his own dream of earning a PhD. I hope he knows that he has accomplished that goal, by the relentless encouragement he has given me and by inspiring me to pursue a PhD. To my parents, I say, thank you for everything! I would also like to thank my dear sister, Dr. Shahnaz Sharmin, for her support, for taking care of my when I got carsick on the school bus, and for teaching me English!

I would also like to thank my former lab mates, Dr. Yi Yang, Dr. Chandra Sekhar Bhagavatula, Dr. Thanapon Noraset, Dr. Michael Lucas, Dr. Zheng Yuan, and soon-to-be Dr. David Demeter. I will always cherish the fun days of random conversations, many on AI, and the help with procrastination. It has been an honor to work with you all. Thank you.

I would like to thank all my professors, especially Dr. Chris Riesbeck. His class on AI Programming using Common Lisp has had an immeasurable impact on how I think as a programmer, without which I would not be able to conduct the experiments that I report on in this dissertation. Special thanks also go out to my favorite school, Northwestern University, for being the great school it is.

Lastly, special thanks to the Starbucks Coffee Company and Red Bull, for obvious reasons. The hundreds of hours I spent sitting at Starbucks doing research and the wings I was given by crates of Red Bull have not escaped my mind. Thank you. You have no idea what you have done for me. (No, I am not funded by either company.)

Dedicated to my three lovely ladies:

My dear wife, the love of my life, Nuzhat Maherin,

and my loving daughters, my princesses, Liyana and Rosie.

Table of Contents

List of Tables

CHAPTER 1

Introduction

Web search and use of social media are two of the most widely used means of accessing information on the World Wide Web. These methods have become ubiquitous, and they greatly influence what information is consumed by Web users. The sources of influence can be as large as a search engine, such as Google, or as small as a person making a prediction on microblogging service, Twitter. Such manners of accessing information influence not only Web users' knowledge, but also their thoughts and beliefs. Search engine results directly dictate what information people view, potentially shaping their opinions about news and facts. On social media, comments and statements that people view as posted by their friends and others determine whom they believe and to what extent, directly or indirectly influencing what beliefs they themselves subscribe to and their own prognostications. The beliefs propagate further, often magnified, by the same mechanisms of information access, disseminating information rapidly and leaving little time for fact checks and provenance determination. This gives rise to a problem: it is not only information that spreads, but also misinformation. The problem is exacerbated by people's tendency to distort the truth, sometimes maliciously, sometimes simply owing to differences in interpretation of information. Due to network effects, misinformation has the potential of far-reaching impact, while magnified wrong beliefs can compound the ill effects of misinformation. The ill effects include, among others, misinformed prognostications that can lead to further undesired effects.

In this dissertation, we address the aforementioned problem by using methods based on our thesis that Web information sources can be harnessed to synthesize accurate predictions of events in an attempt to arrive at the truth. We adopt an approach that acknowledges noise in information and relies on the *wisdom of crowds (WoC)* to derive accurate predictions from data.

The size and the growth of the Web make it challenging to characterize Web information or to separate facts from lies. The result is the shaping of people's thoughts and actions in a way that can be void of truth, and there is little doubt that the aggregate effect of all the information and misinformation on the Web can have significant ramifications, as was witnessed in the snowballing of opinions that culminated in the Tunisian and Egyptian revolutions in 2011 credited for starting the "Arab Spring" in the Middle East. Information propagation on the Web can have real-world outcomes of varying magnitudes, from revolutions to sways in public opinion on elections to the understanding of markets and the economy.

In this circumstance, considerable attention has been given to the "cleanup" of the Web, whereby information can be separated from misinformation, and a certain level of transparency can be attained regarding "true" facts [**1, 2, 3**], as "facts" often include misinformation peddled as such. The quest to discern facts in information has often taken the form of studying bias on the Web, especially of search engines due to their role as a conduit of information for Web users [**4, 5, 6**]. Bias has also been studied as competitive motives of search engines to give preference to their own content as a manner of discriminating against rivals [**7**]. Further, attention has also been given to information credibility of social media [**8, 9, 10**].

The efforts mentioned in the previous paragraph, among others, share the common purpose of providing accountability to statements made online. Viewed in another way, the purpose is to arrive at the truth. One approach the efforts can take is to validate every piece of information and possibly trace its provenance. The approach is susceptible to becoming intractable, as any additional information used to validate a particular piece of information would ideally need to go through the same treatment, resulting in a recursive, fast-growing ontology of information, each node of which would require its own validation.

In contrast to validating every piece of information, we adopt an approach that embraces the presence of misinformation amidst information and seeks to arrive at prognostications from data. In other words, it sidesteps the need to validate every piece of information and instead aims to synthesize accurate predictions from all the available data. To this end, we turn to social media and crowdsourced data in an attempt to extract the truth from it. On the other hand, to assess the credibility of Web search, we study a particular type of bias potentially present in search engine results. The goal of our two-fold approach is to characterize data by understanding its inherent emphasis and to produce predictions that help to surface the truth.

First, we develop methods that result in a system, *PAWS*—Platform for Analyzing Web Search engines—for characterizing the *content emphasis* of Web search engine results: the degree to which differences across search engines' rankings correlate with features of the ranked content, including point of view and advertisements. Using news search results from Google and Bing for a set of query terms spanning different topics, we find no evidence that the engines emphasize results that express positive orientation toward the engine company's products [**11**]. We do find that the engines emphasize particular news sites and that they also favor pages containing their company's advertisements, as opposed to competitor advertisements [**11**]. While measuring search engine bias has become a popular task [**12, 13, 14, 15, 16**], to our knowledge, PAWS is the first system to investigate how relative rankings correlate with important attributes of content including orientation (do search engines favor positive news about their company's products?) and advertisement (do search engines drive traffic to their company's sponsored links?).

Next, we use predictions from social network, Twitter, and financial prediction platform, Estimize, to explore methods to synthesize accurate predictions of events. For this task, we rely on WoC, the phenomenon—first noticed in 1906 [**17**]—that the average of various estimates pertaining to a topic by a crowd of people tends to be more accurate than the estimate by any one individual in the crowd, often matching or even exceeding the accuracy of estimates by experts. The phenomenon has been confirmed [**18**] and replicated [**19, 20, 21, 22**] repeatedly over time, and recently popularized by a bestselling book [**23**] credited for coining the phrase, "wisdom of crowds." We explore the ranges of various features of our data that are optimal for deriving accurate predictions of events while adhering to WoC principles.

Our Twitter data comprises sports predictions for games of particular tournaments, and we explore the principles of WoC to develop methods to use a relatively small *subset* of the predictions to derive accurate predictions of game outcomes. Our Estimize data contains financial predictions—*Revenue* and *Earnings Per Share*—of various companies, and we investigate which of the same tenets of WoC can be leveraged to produce Revenue and Earnings Per Share predictions that rival those by Wall Street, which are also present in the Estimize data. We hypothesize that predictions have the property that they record a person's wager on a future event, allowing verification of the prediction after the event has taken place and providing a measure of the person's predictive ability as well as their emphasis, or a sort of *worth* of their statements.

It is necessary to state that while success in leveraging our hypothesis for accurate predictions of game outcomes or accurate financial predictions can be monetized in the sports betting market or the stock market, neither possibility motivates our work. Instead, we are deeply interested in the scientific inquiry that is at the core of our work. We ask if people's predictions can be aggregated in order to synthesize all their voices into a singular predictive voice. This question is central to our prediction task.

Through the research that we present in this dissertation, we are able to answer our aforementioned question. We find that crowd predictions can indeed be used to arrive at accurate predictions of events (See Sections 4.4 and 5.4). However, we find that the efficacy of our methods is sensitive to the variations in the nature of the data, as evident in our experiments with the sports data from Twitter (See Section 4.3). In our experiments, we use the majority vote by relatively small groups of people to arrive at predictions, and we also use Logistic and Linear Regression models to analyze the interplay of our data features. We find that while some traits of our datasets are readily discernible, our experiments also suggest that there are possibly nuanced relationships among the features that require further study to reveal. Our experiments with the Estimize data further support these observations (See 5.3). Overall, we show the plausibility of our task of synthesizing accurate predictions from crowd predictions using WoC principles.

The remainder of this dissertation is organized as follows. In Chapter 3, we discuss our research on Web search engine content bias. The next two chapters are devoted to our work on predicting events. Specifically, Chapter 4 discusses our work on predicting sporting event outcomes, and Chapter 5 details our work on making financial predictions. All three chapters are organized in sections in the same manner, starting with details of our dataset, followed by a description of our methodology, followed next by a discussion of our experiments and analysis and ending with a summary of our findings.

The contributions of our research are listed below.

1.1. Contributions

The contributions of the research we present in this dissertation can be summarized as follows:

(1) In our research on search engine content bias, we find no evidence in our research on search engine bias that Google or Bing favor results that express positive orientation toward the engine company's products.

(2) We do find that Google tends to favor smaller news outlets, and Bing favors bigger ones.

(3) We also find that Google and Bing rank a page significantly higher than each other when it contains the engine company's advertisements, as opposed to competitor advertisements.

(4) We present our system, PAWS, for the analysis of Web search engine bias.

(5) Our work on predicting events lends credibility to the applicability of crowd wisdom in predicting.

(6) We also show the effectiveness of simple methods such as majority vote, averaging of crowd estimates, and regression (Logistic and Linear).

(7) Our work indicates the presence of signal in crowd predictions, showing that it is not all noise and that algorithmic extraction of signal is possible and has further potential.

CHAPTER 2

Review of Literature

The research centering on the verification of truth in Web information comprises efforts that span a wide gamut of different aspects and objectives pertaining to information. The efforts range from defining the boundaries of neutrality and bias of search engines to measuring the credibility and the veracity of Web information sources to deriving predictions and recommendations from social media. However, they all share the common goal of sifting truth from lies. In this section, we break the literature down by the approaches of the research works as they pertain to the underlying themes of our research: the characterization of Web search engines and the prediction of event outcomes based on the wisdom of crowds (WoC).

2.1. Web Search Engine Content Bias

Characterizing bias on the World Wide Web has long been the subject of research studies, with "neutrality" of the Web being at the center of continual debate. Legal scholars have debated whether search engines should be regulated to ensure neutrality [**24, 25**]. Much research has been conducted on analyzing different aspects of bias.

Chelaru et al. [**12**] have performed sentiment analysis of search engine results, and so have Demartini and Siersdorfer [**13**]. Mowshowitz and Kawaguchi have investigated bias of search engines as evaluated against an "ideal" or "fair" distribution of search results as approximated by the distribution produced by a collection of search engines [**14, 15**].

In [**26**], Kulshrestha et al. propose a generalizable search bias quantification framework that measures the political bias in search results, while Epstein and Robertson find that search engines can influence election outcomes [**27**].

Researchers have observed that there is no “control” engine available to provide a gold standard ranking [**14, 16**]. As a result, bias measurement has remained an open domain of exploration, with researchers striving to gain an understanding of what constitutes neutrality as opposed to bias and to characterize search engines in light of such understanding.

Our goal is not to define or identify neutrality of Web search to compare search results against. Instead, we investigate the tendency of Google and Bing to favor search results that exhibit orientation toward the engine company's products. Notable works that have analyzed search result orientation include [**7**], in which the authors investigate if search engines prioritize their own content over competitors', though they do not find empirical evidence of such behavior, and [**5**], which analyzes content bias in health-related search results.

In addition to studying orientation, we assess the propensity of Google and Bing to prefer certain news outlets to others, finding evidence of such preference. In [**16**], Azzopardi and Owens look for bias in search engines toward news media providers, but do not find any conclusive evidence. Other related works include [**4**], which explores not only content bias in search engines, but also whether search engines drive Web traffic to well-established sites, finding that search engines in China—the engine, Baidu, in particular—tend to drive traffic to well-established sites. Also investigating site preferences of search engines, [**28**] finds that search engines systematically exclude certain sites and certain types of site, though not always intentionally.

We also study search engines' tendency to prioritize advertisements ("ads") for their own vendor company's services over competitors'. Ads are central in driving revenue for search engines. As such, questions have arisen in regards to the role ads play, if any, in the behavior of search engines. The very creators of the Google search engine express, in their seminal paper from around the time of the launch of Google [**29**], that they expect advertising-funded search engines to be inherently biased towards the advertisers and away from the needs of the consumers. [**6**] studies how search engines allocate users across publishers and compete with publishers to attract advertisers. The authors find search engines to be biased against publishers that display many ads. Edelman and Lockwood [**30**] present evidence from as recently as 2011 that search engines give higher rankings to their own services, such as e-mail or maps, than they do to competitors', which our research in [**11**] corroborates.

As of the time of a recent search of the literature, there do not appear to be many studies of ads in relation to bias in search engines apart from the works mentioned in the previous paragraph; most research surrounding ads pertain to other aspects of bias, such as racial bias and discrimination in society as perpetuated by the presence of online ads. As such, our research on Web search engine bias continues to be novel in that it investigates search engine bias *toward* ads that may give the engine's vendor company a commercial advantage.

2.2. Prediction of Events

We approach the task of predicting events by adopting a WoC approach to predict sporting event outcomes using predictions on Twitter and to predict Revenue and Earnings Per Share (EPS) pertaining to various companies' stock releases using predictions on Estimize.

As such, we start by discussing works related to WoC, subsequently reviewing studies pertaining to sports predictions and financial predictions, following which we discuss works that utilize Twitter and Estimize.

2.2.1. Wisdom of Crowds

The WoC phenomenon was originally proposed by Francis Galton in 1907 [**17**]. Since Galton's work, the phenomenon has been studied extensively, reaffirmed by [**18**] and further replicated by [**19, 20, 21, 22**]. Further studies using WoC are surveyed by Lorge et al. in [**31**]. James Surowiecki's 2005 book, *The Wisdom of Crowds* [**23**], popularizes the concept and details its underpinnings, which have since also been studied by [**32**] and [**33**].

More recently, the efficacy of crowds smaller than originally thought required for WoC to work has received considerable attention. Herzog and Hertwig have made the case for fewer judgments [**34**], and Mannes et al. have subsequently found smaller crowds to fare better [**35, 36**] and so have Kao and Couzin [**37**]. [**38**] explores identifying smaller crowds in advance for predictions that beat those from larger crowds.

2.2.2. Sports Predictions

Numerous research works have explored the task of predicting sports, from soccer to cricket to (American) football to basketball.

[**39**] explores predicting English Premier League soccer using 3 months worth of *tweets* (Twitter messages) from Twitter. The work shows that Twitter can indeed be used for predicting games, at least for the dataset it uses. It also shows that when combined with historical data, the Twitter data can lead to higher prediction accuracies than if only the Twitter data or only the historical data is used.

Adopting a unique and interesting approach, the authors of [**40**] bet on the fact that "people's recognition knowledge of names is a proxy for their competitiveness." They posit that people are not familiar with the names of all teams, and they usually hear about a team if the team is good and has a history of performing well. Based on this premise and using data from three soccer tournaments and two tennis tournaments, the authors show that predictions based on their hypothesis are similar to those based on official ranking and that their heuristic performs well when compared with betting odds. The authors conclude in favor of WoC, in that they claim that "aggregating across individual ignorance spawns collective wisdom."

Using team and player features, [**41**] predicts game outcomes for Twenty20 cricket at the English county level using five years of data. The authors show that a simple method such as Naïve Bayes is sufficient to predict game outcomes correctly two-thirds of the time and outperforms gambling benchmarks.

In [**42**], the authors predict outcomes of football games from two seasons of the National Football League (NFL) in the USA. They attempt to predict three types of outcome: winner of game, winner of game *with the spread* (the phrase referring to the point spread that encodes the handicap set for a team by bookmakers), and *over/under* (whether the total points scored in the game will be over or under the over/under line set by bookmakers). While they do not succeed in identifying a strategy that outperforms market efficiency, they find simple features of Twitter data (such as unigrams in tweets pertaining to the home team or the visiting team) to exceed the performance of game statistics features.

The authors of [**43**], in a season-long experiment, find that a crowd of football bettors is systematically biased and performs poorly when predicting which team will win against a point spread. Moreover, the crowd's biases worsen over time. When asked to predict game outcomes by estimating point differentials, however, their predictions are unbiased and wiser, showing the importance of the manner in which predictions are elicited.

[**44**] explores if the sentiment contained in tweets on Twitter can serve as meaningful proxy to predict game outcomes. Additionally, the work inquires if, in the event tweet sentiment does help with predicting game outcomes, the degree of sentiment can predict the magnitude of outcomes. [**45**] looks at predicting games using the latest statuses of the home team and the visiting team.

Our work is distinct from the aforementioned studies, in that we utilize tweets that specifically predict which of two teams will win a game, from which we synthesize predictions for game outcomes. [**46**] also uses predictions on Twitter, but the authors adopt a different approach. While we count tweets *predicting* a team's win toward that team, the authors count tweets *mentioning* a team toward that team, as they consider tweet volume to be a good indicator for team ranking. Additionally, they use the sentiment contained in tweets and score predictions from the tweets as features for their prediction task, for which they find Support Vector Machines to outperform Naïve Bayes and Logistic Regression.

2.2.3. Financial Predictions

In [**47**], the authors seek to compare Artificial Neural Networks with Logistic Regression for the task of predicting Earnings Per Share, finding the former to perform slightly better than Logistic Regression. [**48**] also uses Artificial Neural Networks to make financial predictions – stock prices, to be precise.

In [**49**], the authors adopt a genetic algorithm approach to arrive at stock pick decisions. Using stock votes from a widely used online financial newsletter, the authors use genetic algorithms to identify and rank “experts” in the crowd—in contrast to our WoC approach—whose votes lead to stock pick decisions that are, on average, better than S&P 500 picks for two time periods.

In one study similar to ours [**50**], Wang et al. rely on WoC for financial predictions, but their approach differs from ours in that it combines deep learning and ensemble learning.

[**51**] studies prediction markets that leverage WoC for financial predictions and finds that such markets have “proven to be uncannily accurate in predictions all types of events.” In [**52**], the authors look at using Twitter in combination with WoC, which is similar to our approach with sports predictions, but they use Twitter to source sentiment variations—we source *tweeters'* (Twitter users) predictions—that they combine with collective opinion mining in news articles and financial market movements.

2.2.4. Predictions Using Estimize

The work in [**53**] examines the negative effects of “herding” on Estimize. The work monitors the amount of information a user views before making an earning forecast, finding that the more public information a user views, the less weight they put on their own private information. This has the interesting effect that while it improves the accuracy of each individual forecast, the accuracy of the consensus forecast decreases, since useful private information is prevented from entering the consensus.

[**54**] examines the reliability of online biographies of Estimize users for financial predictions. The authors investigate if the biographical information provided by Estimize contributors are reliable, if the forecast quality is conditional on whether contributors provide their biographical information and names, and if contributors who provide their biographical information but withhold their identities make forecasts with different characteristics than

those who provide their biographical information and identities. They find that contributors who reveal their biographical information are more active on the Estimize platform and issue higher quality forecasts.

In [**55**], the authors examine consensus Revenue and EPS forecasts derived from Estimize and find that they are more accurate than traditional Wall St. equity analysts' consensus forecasts. [**56**] assesses the value of crowdsourced earnings predictions, finding from more than 51,000 predictions from Estimize that the predictions are incrementally useful in forecasting earnings and measuring the market's expectations of earnings.

2.2.5. Predictions Using Twitter

Research works specific to Twitter include systems that recommend tweets [**57**], URLs [**58**], *hashtags* (keyword-based labels popularized by Twitter and now prevalent across the Web) [**59**], and users to follow [**60**]. Additionally, O'Banion and Birnbaum have worked on Twitter-based prediction as well [**61**], to predict voting behavior, and Zaman et al. have worked on predicting information spread on Twitter in [**62**]. Further, [**63**] studies the spread of the practice of astroturfing[1] on Twitter. [**8, 9, 64, 65**] present methods to measure the credibility of content on Twitter. Also focusing on Twitter, Ghosh et al. have tried to find topic experts [**66**], while Sikdar et al. have tried to establish ground truth [**10**].

Our effort differs from the aforementioned works using Twitter in that we predict event outcomes as opposed to making recommendations or predicting behaviors pertaining to Twitter users.

[1] https://en.wikipedia.org/wiki/Astroturfing

CHAPTER 3

Web Search Engine Content Bias

In this chapter, we discuss our research on the content bias in Web search engines, Google and Bing. Our research on measuring content bias of Web search is not premised on arguing that engines should be free of editorial bias. Instead, our goal is to develop methods to *measure* the differences between engine rankings, and provide these measurements to end users. We present PAWS, a Platform for Analyzing Web Search engines. PAWS measures content emphasis: the degree to which differences across search engines' rankings correlate with features of the ranked content [**11**].

To our knowledge PAWS is the first system to investigate how relative rankings correlate with important attributes of content including orientation (do search engines favor positive news about their company's products?) and advertisement (do search engines drive traffic to their company's sponsored links?).

In our research, we do not find evidence that Google and Bing emphasize results that express positive orientation toward the engine company's products. We do find that the engines emphasize particular news sites and that they also favor pages containing their company's advertisements, as opposed to competitor advertisements.

Section 3.1 of this chapter describes the dataset we use for our study. Section 3.2 describes how PAWS gathers search engine results and analyzes search engines for content emphasis. A key challenge faced by PAWS is to identify the orientations of result URLs at scale. To this end, we present a new technique that we discuss in Section 3.2.2 for manually ranking

Manually Selected Queries		
android	artificial intelligence	net neutrality
chrome	aung san suu kyi	open source
gmail	barack obama	republicans
iphone	china	silicon valley
kinect	christian	silicon alley
lumia	conservative	smoking
macbook	democrats	steve jobs
microsoft	gun control	zombies
office	hollywood	
nexus 7	iran	
microsoft	islamic	
surface	liberal	
windows 8	moon landing	

Table 3.1.1. Manually selected queries

results by orientation that minimizes the expected number of human judgments required. Section 3.3 presents PAWS's analysis of content emphasis in news search on Google and Bing followed by a summary of our findings in Section 3.4.

3.1. Dataset

For our study, we use data from news search results, i.e., new articles returned as results for our search queries. News search is an ideal target for analyzing content emphasis, as the results change frequently and often exhibit orientation toward a concept (e.g., good or bad news, reviews, editorials, etc.). Further, for the query terms we use, news links are often returned prominently even on the primary "Web" search pages of Google and Bing.

We use a total of 165 search query terms. 34 of these, shown in Table 3.1.1, are manually selected, chosen to include controversial queries (e.g., religious and political terms) as well as names of popular products, including several products of the engine companies themselves. The product queries are shown in Table 3.3.1. The remaining 131 queries are selected from daily trending queries or "Hot Searches," as reported by Google Trends. We collect the top 10 queries among Hot Searches per day that are not already part of our set of queries. The search results returned for our queries are collected from both Google and Bing as {header, URL, snippet} triples over timeframe, T = 138 days, resulting in the 131 Hot Searches queries and 51,634 unique result URLs. Additionally, HTML source code is collected for every Webpage linked to by those URLs.

3.2. Methodology

One particular challenge for PAWS is to measure content bias in the absence of ground truth, deviation from which would constitute bias. PAWS aims to measure how a search engine's rankings correlate with features of the ranked content. As other researchers have observed, there is no "control" engine available to provide a gold standard ranking [**16, 14**]. Thus, PAWS measures *relative* differences across two primary providers of algorithmic search results today, Google and Bing. PAWS does not explain *why* the differences arise (or which engine is "responsible").

For each pair (q, u) where result URL u is returned by an engine for query q, PAWS calculates a score that indicates whether u tends to be ranked higher for q by Google than by Bing. We refer to the score as $GB(q, u)$, for Google-Bing score. More negative values indicate that Google ranks a result more highly than Bing.

Because the majority of search result clicks occur on the first page of results [**67**], we consider only these results in our experiments. As search results for a query may change over time, we retrieve results for each query once a day, collecting the top 10 results shown by the search engine. As such, the collected search results are balanced across the different queries.

Formally, let $r(d, q, u, e)$ indicate the numeric ranking of each URL u returned on the first page of results for query q on engine e on day d. For URLs u not returned for a given d, q, e, we let $r(d, q, u, e) = \tau$ for a constant τ. Then GB is defined as:

$$GB(q, u) = \sum_{d \in D} r(d, q, u, \text{Google}) - r(d, q, u, \text{Bing}) \tag{3.1}$$

where the sum is computed over dataset D of days d, with each query performed once on both engines each day. GB is computed over only "algorithmic" results, ignoring advertising links on the result page. The constant τ allows GB to account for results returned on the first page by one engine but not the other. In our experiments, we set $\tau = 20$, although altering the parameter by 50% in either direction has negligible impact on our results. In fact, the correlation between the GB scores with $\tau = 20$ and either $\tau = 15$ or $\tau = 25$ is greater than 0.99.

PAWS measures content emphasis by computing the correlation between $GB(q, u)$ and features of the result u. Some features of interest – such as the site u originates from, or whether u contains ads sponsored by the search engine – are relatively straightforward to identify at scale using automated means. However, an additional goal of PAWS is to measure how *orientations* (defined in the next section) in results vary with GB. Below, we discuss why this task is challenging, and present the novel methods PAWS utilizes to perform the task.

A key challenge for PAWS in addition to measuring content bias in the absence of ground truth is to identify the orientations of result URLs at scale. We address this by presenting a new technique for manually ranking results by orientation using dynamic programming to minimize the expected number of human judgments required.

3.2.1. Orientation Acquisition in PAWS

PAWS attempts to measure if $GB(q, u)$ correlates with positive or negative orientation of document u toward query concept q, where orientation is not the valence or relatedness of u to q and is rather the sentiment expressed by u about q. For example, we may ask PAWS if an engine is more likely to show documents reporting good news about a political party or expressing negative views about a product.

Given the large size of the document sets we wish to analyze, automated techniques for detecting orientation would be desirable. Although a variety of related work has been performed on automatic sentiment classification, our task is particularly challenging because a document's orientation toward a product may be buried in a single sentence that differs from the rest of the document's orientation, and sometimes obtaining the orientation requires world knowledge. Using a state-of-the-art sentiment analyzer,[1] we obtained only -0.07 correlation with human ratings on our datasets.

For manual acquisition of orientation labels, crowdsourcing on platforms such as Amazon Mechanical Turk is a typical approach. However, our controlled experiments show that *Workers* (users who complete small tasks) on Amazon Mechanical Turk have difficulty with the task. The responses are of low accuracy even when we ask questions redundantly or restrict to the highest-rated Workers.

Due to the aforementioned challenges, we rely on collecting orientation judgments by labeling instances ourselves. We then validate the labels by computing self-agreement and

[1] http://www.alchemyapi.com/products/features/sentiment-analysis/

inter-annotator agreement with between ourselves on 40 rankings of results for 2 queries, i.e., a total of 380 pairwise comparisons. The Kappa score for self-agreement was 0.617 and for inter-annotator agreement, 0.385. The scores are dramatically better than our sentiment analyzer and Amazon Mechanical Turk baselines, and are considered "fair" agreement, which we believe to be adequate given the subjective nature of our task. The pairwise judgment approach allows for ties in orientation, and produces a partial order of the documents for each query q, avoiding the difficulties of defining a fixed orientation scale.

While this approach requires human judgment, and is, therefore, a manual process, the effort is greatly reduced using a method we develop that uses dynamic programming to minimize the expected number of human judgments required.

3.2.2. Efficient Ranking by Pairwise Judgments

Because the expert judgments that PAWS requires are expensive, we develop a novel approach that ranks documents by orientation while minimizing the expected number of manual judgments. While previous work has considered production of *total* orders from pairwise comparisons (e.g., [**68**]), these are ill-suited to PAWS because orientations are often indistinguishable, i.e., the relative ranking of two search results based on their orientation is subjective, thereby making it near impossible to infer a total order of ranked search results. When the goal is to produce a total ordering of items, Thurstonian models [**69**] can be relied on. In contrast, our goal is to produce a partial ordering of search results that capture the pairwise ordering between results. Further, our focus is on prioritizing which comparison question to ask next to a single annotator, rather than inferring a ranking of items from a set of comparisons, and we assume our annotators to be noiseless. In contrast, Thurstonian models focus on multiple noisy annotators, from whose annotations an order or the probability distribution over an order is inferred. The Bradley-Terry model [**70**] is more related to our task, whereby an annotator is asked to rank several items by pairwise comparison. However, what results from the model is a total ordering as well, which is not a goal we share in our ordering task. With the goal of producing a partial order of search results, we compare our approach to Binary Search and find our method to slightly outperform Binary Search. While a wide variety of work exists in this space and discussion of all of it is out of scope for this minor contribution of the dissertation, our protocol for choosing questions to minimize expected cost in a resource-constrained (i.e., single-annotator) setting of partial (rather than total) orders is novel, to our knowledge.

Formally, we consider placing a new document d at the proper position within a (perhaps empty) partial order O of other documents. Inserting d requires iteratively comparing it to a selected element $i \in O$. If d is of the same orientation as i, the search terminates; otherwise, the search continues in a smaller portion of O, depending on whether d is deemed more positive or more negative than i. Without ties in O, Binary Search is optimal for the insertion task. However, with ties we can sometimes expedite the search by checking larger (i.e., more probable) portions of the partial order first. Our efficient algorithm exploits this intuition.

Let $E_O[J|LB, UB, i]$ indicate the minimum expected number of judgments needed to place a document d within O if we compare first with $i \in O$, given that the position of d is known to lie between lower bound LB and upper bound UB, inclusive. The expression E_O can be decomposed into a sum over the possible outcomes of the comparison of d to i. We compare d to i (one judgment), and if the two are equal in orientation, no additional judgments are required. If they are unequal, we add the minimum expected number of additional judgments required (in terms of E_O), weighted by the probability of each outcome. Thus, E_O can be expressed recursively as:

$$\begin{aligned}(3.2)\quad E_O[J|LB, UB, i] = 1 + &\\ &P(\{LB, .., i-1\}) \min_j E_O[J|LB, i-1, j] + \\ &\qquad P(\{i+1, .., UB\}) \min_j E_O[J|i+1, UB, j]\end{aligned}$$

where $P(S)$ is the probability that the query document d belongs within $S \subseteq O$ in the partial order. In our implementation, we approximate these probabilities using the distribution of documents in O.

At any step of the insertion, computing the comparison element i that minimizes the expected number of judgments (assuming correct responses) is straightforward using dynamic programming and Equation 3.2. We experimentally evaluate our approach, denoted as Min$E[J]$, utilizing several random orderings of the documents we hand-rank in our experiments (see Section 3.3). The results are shown in Table 3.2.1. We allow a variable error rate, where the comparison of documents d_i, d_j belonging to $i, j \in O$ is modeled as a numeric random variable z selected from the density $P(z) \propto e^{-|z-i+j|/\sigma}$, where $z > 1/2$ indicates a $d_i > d_j$ response, $z < 1/2$ indicates $d_i < d_j$, otherwise $d_i = d_j$. So, larger errors are less likely than smaller ones, and the error rate increases with the parameter $\sigma \geq 0$, with $\sigma = 0$ indicating perfect responses. The results show that Min$E[J]$ reduces the average number of judgments required by 6.58% compared to Binary Search. We also find that Min$E[J]$ is slightly more accurate (Table 3.2.1).

		Error Rate			
		0	0.25	0.50	1
Judgments	Binary Search	40.60	41.20	42.73	44.60
	MinE[J]	37.13	38.00	40.07	42.80
Accuracy	Binary Search	1.00	0.99	0.98	0.92
	MinE[J]	1.00	1.00	0.98	0.93

Table 3.2.1. Pairwise judgments and algorithm accuracy. Min$E[J]$ requires 6.58% fewer judgments on average than Binary Search and is slightly more accurate for all error rates.

Query	Extreme GB	Uniform GB
android	-0.53	-0.47
macbook	-0.34	0.1
nexus 7	-0.3	0.16
microsoft office	-0.29	0.11
xbox	-0.15	-0.33
lumia	-0.03	0.36
kinect	-0.01	-0.01
windows 8	0.1	-0.06
chrome	0.3	-0.17
microsoft surface	0.33	-0.32
gmail	0.46	0.21
Avg. Google Products	-0.02	-0.07
Avg. Microsoft Products	-0.01	-0.05

Table 3.3.1. Spearman correlation between GB and orientation rank for product queries. Positive values indicate Google's results favor more positive orientations toward the query. On average, we find no significant evidence of the engines' emphasizing positive orientations toward their company's products.

3.3. Experiments and Analysis

This section discusses the experiments conducted using PAWS to investigate three aspects of result content: orientation toward the engine company's products, presence of the company's advertisements ("ads"), and the site linked to by the result.

3.3.1. Orientation

The first experiment we perform involves orientation measurement on news search results for 11 manually selected product-name queries. From the results for each query, we select

	Google Ads	Microsoft Ads	Facebook Ads	Other Ads
Average over q	-0.02	0.06	0.05	0.01
(Std. Dev.)	(0.08)	(0.04)	(0.06)	(0.06)
Combined	-0.01	0.05	0.06	0.01

Table 3.3.2. Spearman correlation of GB with the presence of ads by the given company. "Average over q" lists the average of 34 correlation values, one for each query. "Combined" lists the correlation when the results from all 34 queries are combined into a single set. When compared against each other, Google and Bing favor content containing their company's own ads, rather than competing ads. The difference between the combined correlation coefficient for Google (-0.01) and that of Microsoft (0.05) and Facebook (0.06) is significant at the $p < 0.001$ level (Fischer r-to-z transformation).

20 results to rank by orientation. We select them in two ways: *Uniform* GB: selection of 20 results approximately uniformly spaced in the set of all results, and *Extreme* GB: selection of 10 results from each of the two ends of the set (i.e., the results most skewed toward being returned by one engine rather than the other). The intuition behind the second set is that the extremal documents are more likely to reflect content emphasis. We rank each set using the manual ranking using Min$E[J]$ as described in the previous section.

The results are shown in Table 3.3.1. While the numbers vary between the two result sets, in neither case does GB show that the engines emphasize positive orientations toward their company's own products. The fact that the average correlations are negative across all queries indicates that Google slightly emphasizes negative results in general, on this dataset.

3.3.2. Advertisements

As our second experiment, we investigate whether the presence of ads in a document linked to by a result URL for an engine or one of its competitors influences the URL's position in search results. Engines have a commercial interest in increasing traffic to their parent company's ads, which makes ads an important content feature to analyze.

We define *ads* broadly to include not only online text and display advertisements, e.g., Google AdSense and Bing Ads, but also links to the search engine's products and services, e.g., YouTube, Google+, etc. To identify ads in each result in our dataset, we manually construct regular expressions for ads by the two major publishers (Google and Microsoft, and their third-party affiliates) and some other companies. We also identify the presence of Facebook *Like* buttons.

We analyze the Spearman correlation between GB and a binary variable indicating the presence of a given company's ads. Table 3.3.2 shows the results. We see that the engines rank a page significantly higher, relatively speaking, when it contains the engine company's ads, as opposed to competitor ads. Compared to Google, Bing also favors content with Facebook Like buttons. The content emphasis on ads seen in this experiment, while not large, may have a non-trivial impact when aggregated over billions of yearly searches.

3.3.3. Sites

The third experiment involves measurements of news search emphasis across different hosts. We show that GB is, in fact, significantly non-uniform for a large number of hosts, indicating that the two engines often prefer different hosts.

Starting with our complete results dataset, we normalize the host names, retaining the suffix. We find 2,990 unique hosts. We focus our analysis on frequent hosts, i.e., those with at least 50 distinct search results in the data.

Of the 150 frequent hosts, 31 have an average GB below 0.35, and 15 have an average GB above 0.65. For frequent hosts, an average GB falling outside of the range [0.35, 0.65] is statistically significant ($p < 0.005$, Monte Carlo simulation with 10,000 trials).

Host	Num. Results	Average Norm. *GB*	Std. Err.
wnd.com	60	0.22	0.02
joystiq.com	117	0.23	0.02
ign.com	202	0.24	0.02
nationalreview.com	76	0.27	0.02
nbcnews.com	179	0.27	0.02
theverge.com	243	0.27	0.01
hollywoodlife.com	198	0.28	0.01
polygon.com	80	0.28	0.02
slate.com	110	0.28	0.02
siliconvalley.com	110	0.29	0.01
...	...	...	...
societyandreligion.com	56	0.69	0.02
cnn.com	851	0.69	0.01
businessweek.com	328	0.70	0.01
upi.com	124	0.72	0.02
itechpost.com	84	0.72	0.02
i4u.com	57	0.74	0.02
ap.org	54	0.74	0.02
msn.com	333	0.75	0.01
betanews.com	95	0.76	0.02
softpedia.com	371	0.77	0.01

Table 3.3.3. Sites emphasis. Google tends to favor smaller news outlets while Bing favors bigger ones.

We see that of the 150 frequent hosts, the 46 (or 31%) discussed above exhibit significantly different ranking behavior in Google than in Bing. Table 3.3.3 lists the 20 hosts at the extremes. We see that Google tends to give relatively higher rank to smaller news sites that may be specialized or politically opinionated (whether conservative or liberal), e.g., wnd.com and slate.com. By contrast, Bing ranks larger US media outlets with a wider print and television news presence relatively higher, e.g., ap.com and cnn.com. Also, Microsoft content (msn.com) and that of its search engine partner (yahoo.com) are shown to rank relatively higher in Bing than in Google in this experiment.

Since our work on PAWS, further research has also explored Web search engine bias [**4, 5, 6, 7**]. However, to the best of our knowledge, our work continues to be the only instance of research to comparatively study search engine content emphasis.

3.4. Summary of Findings

In summary, the findings from our research on search engine content bias are as follow:

(1) We find no evidence in our research on search engine bias that Google or Bing favor results that express positive orientation toward the engine company's products.

(2) We do find that Google tends to favor smaller news outlets, and Bing favors bigger ones.

(3) We also find that Google and Bing rank a page significantly higher than each other when it contains the engine company's ads, as opposed to competitor ads.

CHAPTER 4

Sports Prediction With Twitter

In Chapter 1, we ask if people's predictions can be aggregated in order to synthesize all their voices into a singular predictive voice. One of the research tasks we undertake to answer that question is to explore the possibility of predicting outcomes of sporting events, which we discuss in this chapter. Specifically, we present methods for the prediction of outcomes of games from three tournaments: the *2014 FIFA World Cup* of soccer, the *2015 ICC Cricket World Cup,* and two seasons of *NFL and NCAA* football in the USA.

We explore the possibility and efficacy of predicting game outcomes using tweets on Twitter containing predictions of game outcomes by tweeters. To that end, we rely on the application of the wisdom-of-crowds (WoC) phenomenon. In the WoC setting, we study the use of *majority vote (MV)* aggregation of predictions and compare the approach with Logistic Regression (LR).

We find through experimentation that the predictive voices of a crowd can indeed be synthesized into a singular prediction of an event. In other words, predictions crowdsourced from Twitter do contain signal, and simple methods such as majority vote, averaging of crowd estimates, and LR prove effective for the separation of such signal from noise. Further, our research motivates questions surrounding our task that we slate for future research.

Section 4.1 of this chapter describes the dataset we use for our study. Section 4.2 describes our data features along with the methods we use. Section 4.3 presents our various experiments and associated analysis followed by a summary of our findings in Section 4.4.

4.1. Dataset

For our first study of prediction of events using Web information sources, we use data from microblogging service, Twitter, a social network on the Web where short messages up to 240 characters in length known as *tweets* can be posted by its users. The users are commonly known as *tweeters.* Twitter is a popular platform for people to voice their predictions on about a variety of topics, sports being one of them. As such, and because the platform allows any external computer program to connect to its services via its Application Programming Interface (API), Twitter is a viable source for collecting information in the form of tweets.

We collect two kinds of tweet. Using regular expressions (regexes) handcrafted based on variations of phrases that indicate a sporting team's victory over another—e.g, "Brazil will beat Argentina", "Bangladesh gonna win," "The Vikings will lose," etc.—we collect tweets that predict which of two teams will win a match *(prediction tweet).* We collect such tweets pertaining to three sports tournaments: the 2014 FIFA World Cup (soccer), the 2015 ICC Cricket World Cup, and two seasons of the NFL (football) and the NCAA (college football) from late 2014 to early 2016. We refer to all the prediction tweets collectively as the *predictions dataset,* and refer to tweets pertaining to each tournament as the *soccer dataset,* the *cricket dataset,* and the *football dataset,* respectively, the last collectively referring to the NFL and the NCAA tweets. Details of the datasets are shown in tables in Section 4.3. We also collect random tweets—up to 100 per tweeter—from the same tweeters whose tweets constitute the predictions dataset. The random tweets form a body of text for each tweeter that we subsequently use as their *bag of words (BoW)* in our methods. We refer to this dataset as the *BoW dataset* and the tweets as *BoW tweets.*

Sports Tournament	Games	Predictions	Tweeters	Singletons	Non-singletons
2014 FIFA World Cup	36	32094	22225	14555	7670
2015 ICC Cricket World Cup	46	14378	5588	239	5349
2014-16 NFL and NCAA	173	27602	18127	1503	16624

Table 4.1.1. Predictions dataset. Singletons are tweeters with a single game prediction in the dataset.

The motivation behind using Twitter as the source for predictions is two-fold: firstly, tweets, including those that contain predictions, are abundantly available on Twitter; secondly, the platform's character limit results in tweets that are succinct, allowing for relatively straightforward identification and extraction of predictions without complicated use of Natural Language Processing (NLP) or Understanding (NLU).

4.2. Methodology

Our task pertaining to Twitter is to develop a method to synthesize accurate predictions for the outcome of sporting events *(games)* from subsets of available predictions. The pursuit tests our thesis that Web information can be harnessed to arrive at such predictions and leverages the wisdom of crowds to do so.

Our reliance on a WoC setting motivates the use of groups and informs our methodology. As such, for every game to be predicted, we generate groups of tweeters from our raw data by randomly batching, for each group, an arbitrary number of tweeters who have predicted the outcome of the game.

Therefore, the features of our data we subsequently use with our methods apply to the groups, not to individual tweeters. The methods by which we compute the features are described further into this section.

We begin the rest of this section by describing the models we apply to our data, followed by a discussion of our features.

4.2.1. Models

Implicit in our thesis that Web information can be harnessed to arrive at predictions of event outcomes is the assumption that Web information contain signals that can be methodically extracted for the purpose of prediction. While the task involving our Twitter data is to synthesize event predictions, the goal of the task is to study if features of the data—features devised using hypotheses based on our central thesis—can be leveraged to extract such signals to arrive at accurate predictions.

To study the features of the data, we choose to use a simple approach: to aggregate predictions by generating groups of arbitrary sizes and computing the MV prediction. A prediction tweet predicts which of two sporting teams will win a game, so the MV prediction is the winning-team prediction with the higher number of occurrence among the group predictions.

The human aspect of MV is important to note. Alternate methods include machine learning (ML) models such as *bagging* and *boosting.* With a bagging approach, a number of different classifiers could be trained that would yield separate predictions for a game in an ensemble setting that could then be aggregated to derive a singular prediction for the outcome the game. Similarly, a boosting approach could utilize one of the features discussed in Section 4.2.2, e.g., $Credibility_{grp}$, whereby an aggregate of predictions weighted by the feature could be used toward deriving a prediction for the game outcome. However, we are interested in a WoC setting where the human members of a crowd—the tweeters—are the classifiers whose predictions we aggregate to synthesize a prediction for the game outcome.

Specifically, for a group of predictions, grp, by T tweeters, the $MajorityVote_{grp}$ of the group is given by

$$MajorityVote_{grp} = \underset{team \in Teams}{\operatorname{argmax}} \left(|preds_{team}|\right) \tag{4.1}$$

where $preds_{team}$ denotes predictions that predict a win by team $team \in Teams$, and $Teams = \{team_1, team_2\}$, the set of two teams playing the game.

We consider the MV tantamount to the wisdom of the crowd, where the crowd is the group of tweeters in question. Subsequently, we study the average values of the features pertaining to the groups and how they correlate with the accuracy of predictions and with each other.

We study the features of the data further by using a ML model, for which we choose to use a model that is known to perform well in capturing the relationship among data features while not requiring a very large body of data as is typically required by ML methods such as deep learning. Furthermore, we use handcrafted features based on our hypotheses, instead of relying on the ML model to learn features from the data. Therefore, we choose LR as the ML model to use. An additional advantage of LR as our model of choice is that it is simple enough to make it relatively easier than with some other models to ascertain if findings from the use of the model are attributable to the data or to nuanced artifacts of the model architecture.

4.2.2. Features

We start this section by discussing some of the hypotheses underlying the data features we compute. One of our hypotheses is that people's historic accuracy in predicting events may indicate their future prediction accuracy. Specifically, the more accurate a tweeter's past sports predictions are, the more they can be relied upon for future predictions. To test this hypothesis, we introduce a *credibility* feature that captures the reliability of tweeters' predictions.

We also explore the effects of tweeters' *lemmingness*: the degree to which a tweeter is similar to others in predicting games. The feature is motivated by our hypothesis that *contrarians*—those whose predictions typically oppose the mainstream or herd mentality—may be better predictors. However, we choose to use the more straightforward measure of the *opposite* of lemmingness: the tweeters' tendency to *follow* the crowd, hence the name of the feature.

For our study, we rely on the application of the WoC phenomenon. One of the four requirements for WoC to work as observed by Surowiecki [**23**] is *diversity*, which posits that the estimates by the different individuals in a crowd must be void of systematic bias. In other words, diversity of predictions or estimates is an essential tenet of WoC. Therefore, we use two features pertaining to diversity: the *diversity of predictions* and the *similarity of tweeters* as measured by the similarities of the words they use in their tweets. A related feature that we also use is the *homogeneity* of predictions, which measures the skew of predictions toward the MV prediction.

This section continues with detailed descriptions of our features, and the features are also summarized in Table 4.2.1.

4.2.2.1. Feature $Credibility_{grp}$. Intuitively, a tweeter's past performance in predicting games should indicate how much they can be relied on for future predictions. Additionally, one of the four conditions for the WoC phenomenon to work, as listed by Surowiecki in

Feature	Description
$Credibility_{grp}$	Average accuracy
$Credibility_{max}$	Maximum accuracy
$Lemmingness_{grp}$	Average tendency to agree with others on predictions
$PredictionDiversity_{grp}$	1 – Average Pairwise Cosine Similarity of predictions
$BoWSimilarity_{grp}$	Average Pairwise Cosine Similarity of Bags of Words
$PredictionHomogeneity_{grp}$	Skew of predictions toward majority vote
$NumPredictions_{grp}$	Average number of predictions of tweeters
$NumPredictions_{max}$	Maximum number of predictions of tweeters
$GroupSize_{grp}$	Group size

Table 4.2.1. Features used for Twitter task

[**23**], is *decentralization*. The particular type of decentralization that Surowiecki describes requires that the participants supplying estimates in a WoC setting are able to "specialize and draw on local knowledge." In our Twitter setting, a tweeter's knowledge and understanding of sports constitute their local knowledge, and their specialization informs their degree of expertise.

The notions discussed in the previous paragraph are the motivation for the computation of a tweeter's credibility, $Credibility_t$, based on their prediction history. It is necessary to describe the method by which we determine, at scale, the prediction contained in a tweeter's tweet, which we subsequently validate against the actual outcome of the game their prediction pertains to. As stated in Section 4.1, we utilize handcrafted regexes to identify tweets that contain predictions for games. Due to the simple but stringent structure of our regexes, we expect them to perform well in retrieving relevant tweets using the Twitter API. Nevertheless, it is important to assess the efficacy of our method, which we accomplish by manually inspecting a sample of random tweets for correctness. For every dataset pertaining to the three sports that comprise our domain for the prediction task at hand, we isolate a random sampling of 100 tweets. We then read the tweet to determine the prediction contained in it, thereafter comparing it with the actual outcome of the game it pertains to. Upon such manual comparisons, we find the cricket, soccer, and football datasets to contain, respectively, 17%, 7%, and 8% tweets among the sampled sets containing predictions that contradict the prediction determined computationally. As such, we consider the aforementioned percentages to be the amount of noise in the datasets. Further, considering the sample count of 100 per dataset, we conclude that the sampled sets approximate the distribution of noisy and non-noisy tweets in the entire datasets. We consider the noise percentages to be low enough to consider our datasets informative for our prediction task at hand.

To proceed with feature descriptions, $Credibility_t$, in turn, informs the computation of $Credibility_{grp}$, which measures the credibility of a *group* of tweeters and is the average of the individual credibility measures of every tweeter in the group. Given the short span of time over which the sports tournaments in our prediction dataset occur, we consider it unlikely for a tweeter to improve in their predictive ability over the course of a tournament. As such, our computation of the $Credibility_t$ of a tweeter is agnostic of time: to compute the $Credibility_t$ of a tweeter, we use all their predictions available in the prediction dataset except for the game for which their $Credibility_t$ is to be used to predict its outcome.

$Credibility_t$ is the ratio of the number of correct predictions by a tweeter over all their predictions. For games with a tied outcome, the tweeter is awarded partial credit for their prediction, as the prediction dataset contains only predictions of a team's victory over another and not of tied games.

Specifically, for every tweeter, $t \in T$ tweeters with a prediction for every game, $g \in G$ games in the prediction dataset, their $Credibility_t$ with respect to g is given by

$$Credibility_t = \frac{correct_{g' \in g \setminus G} + \frac{1}{2}\ partial_{g' \in g \setminus G}}{|g \setminus G|} \tag{4.2}$$

where $correct_{g' \in g \setminus G}$ denotes the number of correct predictions by t in all games in G except for game g, and $partial_{g' \in g \setminus G}$ denotes the number of partially-correct predictions for those same games. Additionally, $Credibility_t$ is smoothed as follows using smoothing prior, m, and smoothing strength λ:

$$Smoothed\,Credibility_t = \frac{Credibility_t|G| + \lambda m}{|G| + \lambda} \tag{4.3}$$

Therefore, for a group of predictions, grp, by T tweeters, the $Credibility_{grp}$ of the group is given by

$$Credibility_{grp} = \frac{\sum\limits_{t \in T} Smoothed\,Credibility_t}{|T|} \tag{4.4}$$

The **higher** the $Credibility_{grp}$ of a group of tweeters is, the **more** credible they are considered to be.

4.2.2.2. Feature $Credibility_{max}$. $Credibility_{max}$ is measured using tweeters' individual credibility measures and is simply the maximum of those measures.

Therefore, for group of tweeters, T, using Equation 4.3, $Credibility_{max}$ is given by

$$Credibility_{max} = \max_{t \in T}(Smoothed\,Credibility_t) \tag{4.5}$$

4.2.2.3. Feature ***$Lemmingness_{grp}$*****.** We hypothesize that people who express more contrarian views may exhibit better accuracy in their predictions than those who hold more mainstream views. In other words, people who tend to exhibit more herd mentality—a behavior known to be common to lemmings, as they show a tendency to follow the actions of other lemmings—may be less accurate in their predictions. Therefore, we measure the lemmingness of a group of tweeters, $Lemmingness_{grp}$, by averaging the individual lemmingness of every tweeter in the group, $Lemmingness_t$, which measures the degree to which the tweeter is similar to other tweeters in predicting for the same games. For this measurement, we identify the MV prediction by the other tweeters and the delta between the number of MV predictions and the remaining predictions. If the target tweeter's prediction matches the MV prediction, we award the tweeter with a score equal to the ratio of the delta and the number of total predictions (by the other tweeters); if the target tweeter's prediction differs from the MV prediction, we subtract the same ratio from the tweeter's score. In the case of equally split predictions among the other tweeters, we leave the target tweeter's score unchanged.

Specifically, for every tweeter, $t \in T$ tweeters with a prediction for every game, $g \in G$ games in the prediction dataset, their $Lemmingness_t$ with respect to g is given by

$$Lemmingness_t = \frac{|preds_{match_t}| - |preds_{\neg match_t}|}{|preds_{match_t}| + |preds_{\neg match_t}|} \tag{4.6}$$

where $preds_{match_t}$ denotes the set of predictions by the other tweeters that match the prediction by t, and $preds_{\neg match_t}$ denotes the set of those that do not.

Therefore, for a group of predictions, grp, by T tweeters, the $Lemmingness_{grp}$ of the group is given by

$$Lemmingness_{grp} = \frac{\sum\limits_{t \in T} Lemmingness_t}{|T|} \tag{4.7}$$

The **higher** the $Lemmingness_{grp}$ of a group of tweeters is, the **more** they have agreed with other tweeters in their past common predictions.

One key condition required by the WoC phenomenon is *diversity of opinion*. The notion is that with predictions from a diverse crowd, each individual prediction is likely to be erroneous to some degree, but the errors cancel each other out, resulting in an aggregate that is remarkably close to the truth. As such, we devise several features to measure different aspects of diversity. The features are discussed below.

4.2.2.4. Feature $PredictionDiversity_{grp}$. This feature measures the diversity of a group of tweeters by averaging the individual prediction diversity of each tweeter in the group, $PredictionDiversity_t$: a measure of how different the game predictions by the tweeter are as compared to those by other tweeters in the group. $PredictionDiversity_t$ is computed using the *Bag of Predictions (BoP)* of every tweeter. The BoP is a ternary vector of length equal to the number of games in the prediction dataset. Each slot in the vector corresponds to a game and has a 1 if the tweeter has predicted the first team to win—games are named using the format *FirstTeam-SecondTeam*, and the names are held constant throughout our methodology—a –1 if the tweeter has predicted the second team to win, and a 0 if the tweeter has not predicted for the game. The $PredictionDiversity_t$ of a tweeter is computed as 1 minus the average Pairwise Cosine Similarity between the BoP vectors of the tweeter and all other tweeters, the Cosine Similarity computed pairwise between the tweeter's BoP vector and that of every other tweeter.

Specifically, for every tweeter, $t \in T$ tweeters with a prediction for every game, $g \in G$ games in the prediction dataset and a BoP vector, v_t, their $PredictionDiversity_t$ is given by

$$PredictionDiversity_t = \frac{1 - \sum\limits_{t' \in T'} CosineSimilarity(v_t, v_{t'})}{|T'|} \tag{4.8}$$

where v_t denotes the BoP vector of tweeter t, $v_{t'}$ denotes the BoP vector of every other tweeter $t' \in T$, and $T' = t \setminus T$.

Therefore, for a group of predictions, grp, by T tweeters, the $PredictionDiversity_{grp}$ of the group is given by

$$PredictionDiversity_{grp} = \frac{\sum\limits_{t \in T} PredictionDiversity_t}{|T|} \tag{4.9}$$

The **higher** the $PredictionDiversity_{grp}$ of a group of tweeters is, the **more** diverse the tweeters in the group are in their predictions.

4.2.2.5. Feature $BoWSimilarity_{grp}$**.** This feature measures the diversity of a group of tweeters by averaging a particular *similarity* measure of each tweeter, $BoWSimilarity_t$: how similar the use of words in tweets by the tweeter is to that of other tweeters in the group. It is computed using the BoW of every tweeter. The BoW of a tweeter is a binary vector of length equal to the vocabulary observed in the entire BoW dataset, which is the set of all unique words in all tweets except for a certain set of twenty-five stop words. Each slot in the BoW vector corresponds to a word in the vocabulary and has a 1 if the tweeter has used the word and a 0 otherwise. The $BoWSimilarity_t$ of a tweeter is computed as the average Pairwise Cosine Similarity between the BoW vectors of the tweeters and all other tweeters, the Cosine Similarity computed pairwise between the tweeter's BoW vector and that of every other tweeter.

Specifically, for every tweeter, $t \in T$ tweeters with a prediction for every game, $g \in G$ games in the prediction dataset and a BoW vector, v_t, their $BoWSimilarity_t$ is given by

$$BoWSimilarity_t = \frac{\sum_{t' \in T'} CosineSimilarity(v_t, v_{t'})}{|T'|} \tag{4.10}$$

where v_t denotes the BoW vector of tweeter t, $v_{t'}$ denotes the BoW vector of every other tweeter $t' \in T$, and $T' = t \setminus T$.

Therefore, for a group of predictions, grp, by T tweeters, the $BoWSimilarity_{grp}$ of the group is given by

$$BoWSimilarity_{grp} = \frac{\sum_{t \in T} BoWSimilarity_t}{|T|} \tag{4.11}$$

The **higher** the $BoWSimilarity_{grp}$ of a group of tweeters is, the **more** similar the uses of words in tweets by the tweeters in the group are.

4.2.2.6. Feature $PredictionHomogeneity_{grp}$. For a group of predictions by different tweeters for a particular game, this feature measures the homogeneity of the predictions: how skewed the predictions are toward the MV prediction. It is computed as the ratio between the number of MV predictions to the number of total predictions in the group.

Specifically, for a group of predictions, grp, by T tweeters, the $PredictionHomogeneity_{grp}$ of the group is given by

$$PredictionHomogeneity_{grp} = \frac{max(|preds_{team_1}|, |preds_{team_2}|)}{|preds_{team_1}| + |preds_{team_2}|} \tag{4.12}$$

where $preds_{team_1}$ denotes the set of predictions that predict $team_1$ to win the game, and $preds_{team_2}$ denotes the set of predictions that predict $team_2$ to win.

4.2.2.7. Feature $NumPredictions_{grp}$. This feature counts, for a group of tweeters, the average number of game predictions made by the tweeters in the group. It is computed by averaging the individual game prediction counts of every tweeter in the group, $NumPredictions_t$.

Specifically, for every tweeter, $t \in T$ tweeters with a prediction for every game, $g \in G$ games in the prediction dataset, their $NumPredictions_t$ is given by

$$NumPredictions_t = |G| \tag{4.13}$$

Therefore, for a group of predictions by T tweeters, the $NumPredictions_{grp}$ of the group is given by

$$NumPredictions_{max} = \frac{\sum_{t \in T} NumPredictions_t}{|T|} \tag{4.14}$$

4.2.2.8. Feature $NumPredictions_{max}$**.** This feature counts, for a group of tweeters, the maximum number of game predictions made by the tweeters in the group. It is the maximum value of the individual game prediction counts of every tweeter in the group, $NumPredictions_t$.

Using Equation 4.13, for a group of predictions by T tweeters, the $NumPredictions_{max}$ of the group is given by

$$NumPredictions_{max} = \max_{t \in T}(NumPredictions_t) \tag{4.15}$$

4.2.2.9. Feature $GroupSize_{grp}$**.** For a group of predictions by different tweeters for a particular game, this feature is simply the size of the group of predictions.

Specifically, for a group of predictions, grp, by T tweeters, $GroupSize_{grp}$ of the group is given by

$$GroupSize_{grp} = |grp| \tag{4.16}$$

4.3. Experiments and Analysis

In this section, we discuss experiments performed separately using each of the three sports datasets: the soccer, cricket, and football datasets. Early exploratory experiments

	All Games	Top Games
No. of Games	46	30
No. of Predictions	14378	13701
No. of Tweeters	5588	5559
No. of Singletons	239	511
No. of Non-singletons	5349	5048
Global Singleton Credibility$_t$		0.43
Global Prediction Homogeneity		0.74

Table 4.3.1. Cricket dataset details. The dataset ranges from February 6, 2015 to March 27, 2015. Singletons are tweeters with a single game prediction in the dataset.

show that our prediction methods perform differently for the different datasets, leading us to observe that while all three datasets pertain to sports, the nature of the signals present in the data vary among the particular sports. Therefore, we perform the experiments discussed below separately for the three datasets and report our observations.

4.3.1. Predicting Cricket

4.3.1.1. Majority Vote. We begin with a set of experiments to explore the use of MV to predict cricket game outcomes. First, we predict the winning team for all 46 games in the dataset. We refer to these games as *All Games.* Next, we predict the winning team for only games for which the dataset contains 90 or more predictions. There are 30 such games in the dataset, and we refer to them as the *Top Games.* Table 4.3.1 shows basic statistics pertaining to the two datasets.

Using All Games and the Top Games, for each game, $g \in G$, we generate groups of tweeter predictions of size, $n \in N = \{1, 2, 3, 5, 7, 9\}$. For each group size, n, we select n predictions per game without replacement, repeating the process to yield 10 groups per game. The reason for the repetition is to allow for 10 iterations of prediction per group

Accuracy Metric	All Games			Top Games		
	Average	Max	Min	Average	Max	Min
No. of Correct Predictions	24.88	33	14	18.70	23	12
No. of Partially Correct Predictions	0.95	1	0	0.92	1	0
No. of Wrong Predictions	15.87	31	8	10.38	17	6
No. of Total Predictions	41.70	46	30	30.00	30	30
Percentage of Correct Predictions (%)	60	75	30	62	77	40

Table 4.3.2. Cricket prediction accuracy. Counts shown for all 46 games on the left (*All Games* with 2,502 predictions) and the top 30 games on the right (*Top Games:* those with 90 or more predictions, totaling 1,800 predictions). Averages are computed over 60 iterations.

size across all the games in the dataset. However, when All Games are used, for the set of 16 games not among the Top Games, certain larger group sizes limit the number of groups to fewer than 10, resulting in those particular games being omitted from the prediction iterations pertaining to the responsible group sizes. With the Top Games, however, given the range of N and the prediction count threshold of 90, every game yields 10 groups. With 10 iterations of prediction per group size, we have a total of 60 iterations of prediction per dataset.

Utilizing each group for MV prediction (the winner of the game) using Equation 4.1, we make 2,502 predictions for All Games and 1,800 predictions for the Top Games.

We evaluate the accuracy of the predictions by counting the number of correct predictions, the number of partially correct predictions (for tied games that tweeters predicted one team would win), and the number of wrong predictions. These accuracies are shown in Table 4.3.2.

As mentioned before, predicting with different sports datasets is observed in exploratory experiments to yield different levels of performance, which leads us to assume such differences are manifest due to differences in the tweeter predictions for individual games. As such, it is

Feature	All Games			Top Games		
	Average	Max	Min	Average	Max	Min
$\text{Credibility}_{\text{grp}}$	0.48	0.54	0.35	0.54	0.54	0.53
$\text{Lemmingness}_{\text{grp}}$	0.27	0.50	0.06	0.34	0.53	0.18
$\text{PredictionDiversity}_{\text{grp}}$	0.64	0.95	0	0.7	0.95	0
$\text{BoWSimilarity}_{\text{grp}}$	0.21	0.36	0	0.23	0.38	0

Table 4.3.3. Cricket feature value average. The averages are computed over 60 iterations, each yielding an average over All Games on the left and the Top Games on the right.

important to have enough tweeter predictions per game in order to capture the characteristics of predictions pertaining to that particular game, which is the motivation for the Top Games. Selecting 90 predictions as the minimum threshold results in a sizable set of games to predict for, hence the choice of 90.

The percentages on the bottom line of Table 4.3.2 appear to corroborate the assumption that having more tweeter predictions per game is likely to yield better MV accuracy. To study the effect further, we look at the average values of the features pertaining to the tweeter predictions. The averages are first computed over the games, the resulting averages then averaged once more over the 60 iterations.

We observe that for both the All Games and the Top Games datasets, the value of $\text{Credibility}_{\text{grp}}$ hovers around 0.50, which indicates a lack of predicting expertise among tweeters when the credibility of their predictions is measured solely based on their history of predictions. However, in the limited dataset of the Top Games, the feature has a higher value. A possible explanation for such gain, besides the possibility that it is so because of noise, is the presence of relatively more tweets per game in the dataset than in the All Games dataset when considered in the context of a noticeably low global $\text{Credibility}_{\text{grp}}$ of *singletons:* tweeters who have a single tweet in the prediction dataset. While a tweeter's credibility, $\text{Credibility}_{\text{t}}$, is otherwise computed based on their prediction history, a singleton's credibility is computed based on the overall prediction performance of all singletons in the dataset. Because of the higher number of tweets and the distribution of non-singletons between the two datasets, the Top Games contain more predictions per non-singleton, thereby likely diverging their $\text{Credibility}_{\text{grp}}$ from the global singleton $\text{Credibility}_{\text{t}}$ toward a more realistic value.

We further notice that the value of $\text{Lemmingness}_{\text{grp}}$ is very low, indicating that tweeters have a tendency to disagree with the majority vote when predicting when measured across games—they are more contrarian. It may appear paradoxical, as most tweeters' disagreeing with one prediction essentially forms a majority vote in favor of the other prediction. However, it is still plausible that a significant number of tweeters disagree with the majority vote in enough games for this effect to be observed in aggregate across all their predictions for various games. This is an effect we first notice in our preliminary explorations of the data and warrants further investigation.

We acknowledge that the aforementioned effect related to $\text{Lemmingness}_{\text{grp}}$ may be the reason for our next observation that the prediction diversity of tweeters across either dataset is considerably high, as evident in the value of $\text{PredictionDiversity}_{\text{grp}}$. It could be argued that for tweeters to disagree with the majority often, yet not form their own majority vote, they would have to have be diverse among themselves in their predictions, i.e., sometimes agree with other tweeters with whom they co-predict games and, at other times, disagree. We also note that value of $\text{PredictionDiversity}_{\text{grp}}$ is noticeably high for both datasets, which would be commensurate with the diversity requirement of the WoC phenomenon. We further note that the delta in $\text{PredictionDiversity}_{\text{grp}}$ between the two datasets correlates with the delta in the percentage of correct predictions in that both have higher values for the Top Games than for All Games.

Feature Pair	All Games		Top Games	
	Coeff.	p-value	Coeff.	p-value
Correct predictions : $\text{Credibility}_{\text{grp}}$	0.42	0	0	1
Correct predictions : $\text{Lemmingness}_{\text{grp}}$	0.23	0.09	0.13	0.32
Correct predictions : $\text{PredictionDiversity}_{\text{grp}}$	-0.17	0.20	0.03	0.81
Correct predictions : $\text{BoWSimilarity}_{\text{grp}}$	-0.17	0.18	-0.08	0.53
$\text{Credibility}_{\text{grp}}$: $\text{Lemmingness}_{\text{grp}}$	0.33	0.01	0.29	0.03
$\text{Credibility}_{\text{grp}}$: $\text{PredictionDiversity}_{\text{grp}}$	-0.13	0.31	0.01	0.95
$\text{Credibility}_{\text{grp}}$: $\text{BoWSimilarity}_{\text{grp}}$	-0.07	0.58	-0.03	0.84
$\text{Lemmingness}_{\text{grp}}$: $\text{PredictionDiversity}_{\text{grp}}$	0.03	0.81	-0.06	0.67
$\text{Lemmingness}_{\text{grp}}$: $\text{BoWSimilarity}_{\text{grp}}$	0.07	0.58	-0.13	0.32
$\text{PredictionDiversity}_{\text{grp}}$: $\text{BoWSimilarity}_{\text{grp}}$	0.97	0	0.94	0

Table 4.3.4. Cricket feature pair correlation. *Coeff.* is Pearson's Correlation Coefficient, computed over 60 iterations.

The observations discussed thus far about the features do not readily indicate any influence of the features, if any, on the number of correct predictions. To study the relationship between the correct predictions and the features further, we compute pairwise Pearson's Correlation Coefficient ("correlation") along with the related p-value between the number of correct predictions and every feature. We also compute the pairwise correlation between every feature pair to gain further insight into the relationship among the features.

Per Table 4.3.4, $\text{Credibility}_{\text{grp}}$ has fairly strong correlation with the number of correct predictions, and the correlation is statistically significant with $p < 0.05$. The extreme difference in values of the correlation coefficient and the p-value between All Games and the Top Games, however, is intriguing for the feature pair. Turning to Table 4.3.3 for insight, we notice that the values of $\text{Credibility}_{\text{grp}}$ have a small range, which may explain the intriguing effect. However, we slate such effects as warranting further study in future investigations.

Also noteworthy is the correlation between $\text{Lemmingness}_{\text{grp}}$ and the number of correct predictions. Though shy of being statistically significant, the correlation contradicts the possibility of less lemmingness (more contrarianness) of tweeters leading to more accurate predictions. The effect of lemmingness is more pronounced in the correlation between $\text{Lemmingness}_{\text{grp}}$ and $\text{Credibility}_{\text{grp}}$, which suggests that the more lemming a tweeter is, the more accurate their predictions are, which lends support to the aforementioned correlation between lemmingness and the number of correct predictions.

Finally, we note that $\text{PredictionDiversity}_{\text{grp}}$ and $\text{BoWSimilarity}_{\text{grp}}$ show an extraordinarily strong and statistically significant correlation. The suggestion implicit in the correlation is counterintuitive, as it suggest that tweeters with similar use of vocabulary in their tweets tend to predict differently. While that is not implausible, the strong correlation is surprising and warrants further study.

4.3.1.2. Logistic Regression. In addition to performing MV experiments, we perform LR to further study our data features and their relationships and to utilize them to predict game outcomes. The features used for LR are listed in Table 4.2.1.

Similarly to our approach with the MV experiments, we perform the LR experiments first on All Games, then on the Top Games. For each game in either dataset, we use LR to predict game outcomes using the same groups of prediction that we use to predict game outcomes in the MV experiments. The feature values pertaining to each group are averaged to form a row in the data used for LR, thereby generating multiple rows per game, each corresponding to a prediction to be made by LR.

It is important to understand one further detail about the LR experiments. The LR predictions are made in the manner of *cross validation*, in that a separate LR model is trained to predict every game, for which the game is held out from the training data. In other words, each LR model is trained on the grouped data mentioned in the previous

	All Games			Top Games	
	LR	MV		LR	MV
Average	0.67	0.61	Average	0.66	0.61

Table 4.3.5. Cricket Logistic Regression (LR) results: Average accuracies. LR accuracy averages are compared to Majority Vote (MV) accuracy averages.

paragraph comprising all games except the one to be predicted for. Specifically, for every game $g \in G$, predictions about the outcome of g are held out, and the LR model is trained on all predictions pertaining to every game $g \setminus G$. The resulting LR models are later used to predict game outcomes during the LR experiments.

Table 4.3.5 lists the LR accuracies. Also listed in the same table are MV accuracies for predictions made using *all* the predictions in the prediction dataset for a game (*whole-set MV predictions,* for the purpose of this discussion). In other words, while the MV experiments use the same number of predictions per group as used for the LR experiments, the whole-set MV predictions utilize the entire crowd of tweeters whose predictions are available per game.

We observe that LR performs better than MV for both All Games and the Top Games. For a more detailed view, Table 4.3.6 lists the top five LR accuracies along with their MV counterparts and corresponding averages. On average, LR is 7.78% more accurate for All Games and 12.64% more accurate for the Top Games. When compared against the top five MV accuracies, the corresponding LR accuracies are respectively 1.08% and 6.59% more accurate on average for the two datasets, as shown in Table 4.3.7. The complete table (Table 7.0.1) is available in the Appendix.

Game	All Games		Game	Top Games	
	LR	MV		LR	MV
bangladesh-scotland	1	0.98	new zealand-scotland	1	0.91
pakistan-ireland	0.97	0.85	india-zimbabwe	0.98	0.91
new zealand-afghanistan	0.97	0.90	australia-scotland	0.98	0.92
india-zimbabwe	0.97	0.87	pakistan-zimbabwe	0.97	0.75
australia-scotland	0.97	0.92	pakistan-united arab emirates	0.97	0.88
Average	0.97	0.90	Average	0.98	0.87

Table 4.3.6. Cricket Logistic Regression (LR) results: Top 5 LR accuracies. The top 5 LR accuracies are compared to the corresponding Majority Vote (MV) accuracies.

Game	All Games		Game	Top Games	
	LR	MV		LR	MV
bangladesh-scotland	1	0.98	australia-scotland	0.98	0.92
australia-scotland	0.97	0.92	new zealand-scotland	1	0.91
new zealand-afghanistan	0.97	0.90	india-zimbabwe	0.98	0.91
new zealand-scotland	0.95	0.95	south africa-pakistan	0.97	0.90
west indies-ireland	0.82	0.90	new zealand-afghanistan	0.92	0.89
Average	0.94	0.93	Average	0.97	0.91

Table 4.3.7. Cricket Logistic Regression (LR) results: Top 5 MV accuracies. The top 5 Majority Vote (MV) accuracies are compared to the top 5 LR accuracies.

4.3.2. Predicting Soccer

4.3.2.1. Majority Vote. To predict outcomes of soccer games, we follow the same structure of experiments as with predicting for cricket. First, we predict the winning team for all 36 games in the dataset or All Games. Next, we predict the winning team for only games for which the dataset contains 90 or more predictions, the Top Games. There are 30 such games in the dataset. Table 4.3.8 shows basic statistics pertaining to the two datasets.

Using the Soccer dataset, we make 2,013 predictions for All Games and 1,800 predictions for the Top Games. Prediction accuracies are shown in Table 4.3.9.

	All Games	Top Games
No. of Games	36	30
No. of Predictions	32094	31928
No. of Tweeters	22225	22211
No. of Singletons	14555	14639
No. of Non-singletons	7670	7572
Global Singleton Credibility$_{\mathrm{t}}$		0.43
Global Prediction Homogeneity		0.73

Table 4.3.8. Soccer dataset details. The dataset ranges from June 13, 2014 to June 26, 2014. Singletons are tweeters with a single game prediction in the dataset.

Accuracy Metric	All Games			Top Games		
	Average	Max	Min	Average	Max	Min
No. of Correct Predictions	13.72	20	9	11.10	16	7
No. of Partially Correct Predictions	8.38	9	2	8.40	9	2
No. of Wrong Predictions	11.45	20	4	10.50	17	5
No. of Total Predictions	33.55	36	30	30.00	30	30
Percentage of Correct Predictions (%)	41	61	26	37	53	23

Table 4.3.9. Soccer prediction accuracy. Counts shown for all 36 games on the left (*All Games* with 2,013 predictions) and the top 30 games on the right (*Top Games:* those with 90 or more predictions, totaling 1,800 predictions). Averages are computed over 60 iterations.

MV prediction accuracy is not remarkable for either All Games or the Top Games. In fact, the very low accuracy of MV prediction for the Top Games may appear to indicate that minority vote is more indicative of correct predictions. However, further analysis is warranted before arriving at a conclusion. As such, we look at the average values of the features pertaining to the tweeter predictions.

We note that an extremely low values of Lemmingness$_{\mathrm{grp}}$, indicating the tweeters are very contrarian (See discussion in Section 4.3.1), may be the reason behind the considerably large PredictionDiversity$_{\mathrm{grp}}$ values.

Feature	All Games			Top Games		
	Average	Max	Min	Average	Max	Min
$\text{Credibility}_{\text{grp}}$	0.45	0.49	0.40	0.48	0.49	0.48
$\text{Lemmingness}_{\text{grp}}$	0.03	0.10	0	0.04	0.12	0
$\text{PredictionDiversity}_{\text{grp}}$	0.72	1	0	0.77	1	0
$\text{BoWSimilarity}_{\text{grp}}$	0.26	0.44	0	0.29	0.42	0

Table 4.3.10. Soccer feature value average. The averages are computed over 60 iterations.

We observe that for both the All Games and the Top Games datasets, the value of $\text{Credibility}_{\text{grp}}$, with values near 0.50, indicate a lack of predicting expertise among tweeters when the credibility of their predictions is measured solely based on their history of predictions.

In Table 4.3.11, $\text{Credibility}_{\text{grp}}$ shows a much larger correlation with the number of correct predictions for All Games than does any other feature, and it it also statistically significant with a $p < 0.05$. Therefore, the prediction accuracies in Table 4.3.9 may be assumed to be mostly driven by $\text{Credibility}_{\text{grp}}$.

We also notice a noticeable but negative correlation between $\text{Lemmingness}_{\text{grp}}$ and the number of correct predictions, indicating that the more contrarian the tweeters are, the more accurate they likely are. The observation lends further support, though weakly, to the notion of contrarians being better predictors and may have also contributed to the number of correct predictions.

The surprisingly strong correlation we observe between $\text{PredictionDiversity}_{\text{grp}}$ and $\text{BoWSimilarity}_{\text{grp}}$ for cricket is also observed for soccer, and may require a deeper exploration of why such may be the case for predictions on Twitter.

Feature Pair	All Games		Top Games	
	Coeff.	p-value	Coeff.	p-value
Correct predictions : $\text{Credibility}_{\text{grp}}$	0.31	0.02	-0.04	0.77
Correct predictions : $\text{Lemmingness}_{\text{grp}}$	-0.18	0.18	-0.02	0.88
Correct predictions : $\text{PredictionDiversity}_{\text{grp}}$	-0.12	0.35	-0.22	0.09
Correct predictions : $\text{BoWSimilarity}_{\text{grp}}$	-0.08	0.57	-0.26	0.05
$\text{Credibility}_{\text{grp}}$: $\text{Lemmingness}_{\text{grp}}$	0.06	0.63	-0.10	0.46
$\text{Credibility}_{\text{grp}}$: $\text{PredictionDiversity}_{\text{grp}}$	-0.30	0.02	-0.13	0.32
$\text{Credibility}_{\text{grp}}$: $\text{BoWSimilarity}_{\text{grp}}$	-0.32	0.01	-0.18	0.17
$\text{Lemmingness}_{\text{grp}}$: $\text{PredictionDiversity}_{\text{grp}}$	-0.11	0.41	-0.25	0.05
$\text{Lemmingness}_{\text{grp}}$: $\text{BoWSimilarity}_{\text{grp}}$	-0.12	0.36	-0.26	0.04
$\text{PredictionDiversity}_{\text{grp}}$: $\text{BoWSimilarity}_{\text{grp}}$	0.96	0	0.96	0

Table 4.3.11. Soccer feature pair correlation. *Coeff.* is Pearson's Correlation Coefficient, computed over 60 iterations.

	All Games			Top Games	
	LR	MV		LR	MV
Average	0.61	0.57	Average	0.49	0.52

Table 4.3.12. Soccer Logistic Regression (LR) results: Average accuracies. LR accuracy averages are compared to Majority Vote (MV) accuracy averages.

4.3.2.2. Logistic Regression. As with the other two sports datasets, in addition to performing MV experiments, we perform LR to further study our soccer data features and their relationships and to utilize them to predict game outcomes. The features used for LR are listed in Table 4.2.1.

Table 4.3.12 lists the LR accuracies. Also listed in the same table are accuracies for whole-set MV predictions. An important point to note about the results is that it omits tied games, as our chosen methodology used for counting predictions by the LR model lends an unfair advantage to the model over MV predictions.

The LR accuracies offer interesting insights into sports prediction. We see in Table 4.3.13 that the top five LR accuracies are significant higher than their MV counterparts for both

Game	All Games		Game	Top Games	
	LR	MV		LR	MV
mexico-cameroon	0.90	0.92	spain-netherlands	0.82	0.22
japan-colombia	0.88	0.82	belgium-russia	0.77	0.25
italy-uruguay	0.79	0.42	ghana-usa	0.77	0.44
ghana-usa	0.73	0.41	usa-germany	0.75	0.36
spain-netherlands	0.73	0.28	nigeria-argentina	0.67	0.37
Average	0.81	0.57	Average	0.75	0.33

Table 4.3.13. Soccer Logistic Regression (LR) results: Top 5 LR accuracies. The top 5 LR accuracies are compared to the corresponding Majority Vote (MV) accuracies.

All Games and the Top Games. However, as Table 4.3.14 shows, LR accuracy is significantly worse than MV accuracy when compared with the top five MV accuracies. We recall the scenario with cricket, where LR predictions have higher accuracy regardless of which end of the accuracy spectrum they are compared with MV accuracies in. Soccer presents a different scenario, as more evident in the complete table of accuracies (Table 7.0.2, tied games not shown) shown in the Appendix. Firstly, we observe that LR accuracies for All Games and the Top Games lag behind the corresponding MV accuracies with average accuracy of 0.47 vs. 0.47 and 0.49 vs. 0.52, respectively. In other words, the higher-than-MV performances demonstrated by LR are limited to certain games. Secondly, we observe that the accuracies are *strongly inversely* correlated with a correlation coefficient of -0.87. We hypothesize that the effects we discuss in this paragraph may be a result of the nature of the sport, how people predict soccer games, and the occasional “upsets,” whereby a team majority-voted to win unexpectedly loses the game. The effects are fascinating, one that we consider a good candidate for future research.

Game	All Games		Game	Top Games	
	LR	MV		LR	MV
mexico-cameroon	0.90	0.92	belgium-algeria	0.05	0.89
belgium-algeria	0.65	0.85	portugal-ghana	0.1	0.76
portugal-ghana	0.70	0.83	switzerland-france	0.08	0.75
japan-colombia	0.88	0.82	germany-portugal	0.35	0.61
switzerland-france	0.67	0.78	korea-algeria	0.33	0.58
Average	0.76	0.84	Average	0.18	0.72

Table 4.3.14. Soccer Logistic Regression (LR) results: Top 5 MV accuracies. The top 5 Majority Vote (MV) accuracies are compared to the corresponding LR accuracies.

4.3.3. Predicting Football

4.3.3.1. Majority Vote. To predict outcomes of football games, we follow the same structure of experiments as with predicting for cricket and for soccer. However, we omit 3 games from all our predictions, as the games appear in the dataset from both football seasons, thereby comprising predictions that are not from the same timeframe. Therefore, first, we predict the winning team for all 170 games in the dataset or All Games. Next, we predict the winning team for only games for which the dataset contains 90 or more predictions, the Top Games. There are 9 such games in the dataset after omitting the 3 aforementioned games. Table 4.3.15 shows basic statistics pertaining to the two datasets.

Using the Football dataset, we make 3,416 predictions for All Games and 540 predictions for the Top Games. Prediction accuracies are shown in Table 4.3.16.

MV prediction accuracy is not remarkable for either All Games or the Top Games. Looking at the average values of the features pertaining to the tweeter predictions, we note that PredictionDiversity$_{\text{grp}}$ is considerably high among Top Games predictions. A

	All Games	Top Games
No. of Games	173	12
No. of Predictions	27602	25837
No. of Tweeters	18127	11445
No. of Singletons	1503	4740
No. of Non-singletons	16624	6705
Global Singleton Credibility$_{\mathrm{t}}$		0.375
Global Prediction Homogeneity		0.698

Table 4.3.15. Football dataset details. The dataset ranges over two NFL and NCAA seasons spanning from December 18, 2014 to February 6, 2016. Singletons are tweeters with a single game prediction in the dataset.

Accuracy Metric	All Games			Top Games		
	Average	Max	Min	Average	Max	Min
No. of Correct Predictions	22.78	95	1	1.63	4	0
No. of Wrong Predictions	34.15	112	6	7.37	9	5
No. of Total Predictions	56.93	170	9	9.00	9	9
Percentage of Correct Predictions (%)	36	56	11	18	44	0

Table 4.3.16. Football prediction accuracy. Counts shown for all 170 games on the left (*All Games* with 3416 predictions) and the top 9 games on the right (*Top Games:* those with 90 or more predictions, totaling 540 predictions). Averages are computed over 60 iterations.

PredictionDiversity$_{\mathrm{max}}$ of 0.85 suggests that there are sets of tweeters who co-predicted certain games and were very diverse within their respective groups. This raises the question of whether diversity results in high accuracy of predictions for football.

4.3.3.2. Logistic Regression. Tables 4.3.18 and 4.3.19 do show some very high LR accuracies for both All Games and the Top Games, and in both tables. However, the feature correlations in Table 4.3.20 raise questions about noise in the data.

We observe significantly strong correlation between the number of correct predictions and both Credibility$_{\mathrm{grp}}$ and Lemmingness$_{\mathrm{grp}}$. Further, a group's lemmingness has perfect

Feature	All Games			Top Games		
	Average	Max	Min	Average	Max	Min
$\text{Credibility}_{\text{grp}}$	0.12	0.36	0.02	0.27	0.28	0.27
$\text{Lemmingness}_{\text{grp}}$	0.19	0.70	0	0.10	0.28	-0.05
$\text{PredictionDiversity}_{\text{grp}}$	0.20	0.87	0	0.50	0.85	0
$\text{BoWSimilarity}_{\text{grp}}$	0.12	0.51	0	0.35	0.57	0

Table 4.3.17. Football feature value average. The averages are computed over 60 iterations.

Game	All Games		Game	Top Games	
	LR	MV		LR	MV
vikings-49ers	1	0	lions-cowboys	1	0.08
titans-browns	1	0	ravens-patriots	1	0.09
rams-vikings	1	0	seahawks-packers	1	0.05
dolphins-bills	1	0	packers-cowboys	0.98	0.22
dolphins-jets	1	0	cardinals-panthers	0.98	0.01
Average	1	0	Average	0.99	0.09

Table 4.3.18. Football Logistic Regression (LR) results: Top 5 LR accuracies. The top 5 LR accuracies are compared to the corresponding Majority Vote (MV) accuracies.

correlation with their credibility, as does their prediction diversity with their use of vocabulary. Reflecting on such observations, we conclude that such surprising correlations are likely merely artifacts of the lack of predictions across most games, as LR is trained on 170 games for All Games, all but 12 of which has fewer than 90 prediction tweets. In other words, we are led to consider the training data for LR noisy.

Considering the possibility of noise in the data, yet observing high LR accuracies for certain games, it not readily discernible if the high LR accuracies when MV predictions lag so far behind are strokes of luck given perhaps unexpected game "upsets" or if LR is able to actually learn to detect signal given the prediction-rich Top Games. We note that when

Game	All Games		Game	Top Games	
	LR	MV		LR	MV
bengals-bills	1	1	florida state-oregon	0.33	0.52
buccaneers-falcons	1	1	ohio state-alabama	0.37	0.51
rams-bengals	1	1	seahawks-patriots	0.65	0.5
jaguars-ravens	0.33	1	bengals-colts	0.85	0.26
colts-falcons	0.29	1	packers-cowboys	0.98	0.22
Average	0.72	1	Average	0.64	0.40

Table 4.3.19. Football Logistic Regression (LR) results: Top 5 MV accuracies. The top 5 Majority Vote (MV) accuracies are compared to the corresponding LR accuracies.

Feature Pair	All Games		Top Games	
	Coeff.	p-value	Coeff.	p-value
Correct predictions : $\text{Credibility}_{\text{grp}}$	0.95	0	-0.06	0.66
Correct predictions : $\text{Lemmingness}_{\text{grp}}$	0.95	0	0.17	0.21
Correct predictions : $\text{PredictionDiversity}_{\text{grp}}$	0.29	0.02	-0.22	0.10
Correct predictions : $\text{BoWSimilarity}_{\text{grp}}$	0.29	0.03	-0.19	0.15
$\text{Credibility}_{\text{grp}}$: $\text{Lemmingness}_{\text{grp}}$	1	0	-0.54	0
$\text{Credibility}_{\text{grp}}$: $\text{PredictionDiversity}_{\text{grp}}$	0.46	0	-0.13	0.34
$\text{Credibility}_{\text{grp}}$: $\text{BoWSimilarity}_{\text{grp}}$	0.45	0	-0.09	0.48
$\text{Lemmingness}_{\text{grp}}$: $\text{PredictionDiversity}_{\text{grp}}$	0.44	0	-0.12	0.36
$\text{Lemmingness}_{\text{grp}}$: $\text{BoWSimilarity}_{\text{grp}}$	0.44	0	-0.06	0.64
$\text{PredictionDiversity}_{\text{grp}}$: $\text{BoWSimilarity}_{\text{grp}}$	1	0	0.93	0

Table 4.3.20. Football feature pair correlation. *Coeff.* is Pearson's Correlation Coefficient, computed over 60 iterations.

evaluated based on the top five Top Games, LR has robust performance whether the games are sorted by LR accuracy or MV accuracy.

We conclude that the observations with football predictions show that sports domains indeed have their own characteristics and quirks when looked at through the prediction task at hand. It is probably fair to assume that additional features that capture the characteristics of the teams and further traits of the tweeters are likely to yield more insight into how to predict games given a sport. We find it difficult to dismiss the observations as simply

	All Games			Top Games	
	LR	MV		LR	MV
Average	0.54	0.50	Average	0.8	0.25

Table 4.3.21. Football Logistic Regression (LR) results: Average accuracies. LR accuracy averages are compared to Majority Vote (MV) accuracy averages.

artifacts of noise data, and are left encouraged to slate future research to investigate these observations and the interactions of features in further detail.

The surprising strong correlation we observe between $\text{PredictionDiversity}_{\text{grp}}$ and $\text{BoWSimilarity}_{\text{grp}}$ for cricket is also observed for football, and may require a deeper exploration of why such may be the case for predictions on Twitter.

4.4. Summary of Findings

In summary, the findings from our research on predicting events are as follow:

(1) The results of our experiments lend credibility to the applicability of crowd wisdom in predicting. We find that the predictive voices of a crowd can indeed be synthesized into a singular prediction of an event.

(2) Our experiments also suggest that there are possibly nuanced relationships among our data features, as we observe in their pairwise correlations, that require further study to reveal.

(3) We also observe that LR and MV exhibit disparate performances for the Top Games in soccer, and that there is stark contrast between their performances for All Games and the Top Games. Are these effects attributable to our methods, or could they be due to the finicky nature of the sport? The Top Games have more tweets per game, indicating popularity. Could unexpected and surprising game outcomes then

have a role to play whereby popular predictions are proven wrong? We slate such inquiries for future research.

(4) The role of credibility continues to show importance, yet it is difficult to confirm. It warrants further study as well.

(5) Contrarianness and diversity appear salient in more than one occasion and deserve further study rather than being discredited due to the overall nature of the dataset.

(6) Simple methods such as majority vote, averaging of crowd estimates, and LR prove notably effective for our tasks.

(7) Our work indicates the presence of signal in crowd predictions, showing that it is not all noise and that algorithmic extraction of signal is possible and has further potential.

CHAPTER 5

Financial Prediction With Estimize

Motivated by the inspiring results of our experiments with predicting sporting even outcomes, we study the application of similar approaches to financial predictions. Continuing to rely on the wisdom-of-crowds (WoC) phenomenon, we turn to predictions made on Estimize that we aggregate using majority vote and Linear Regression to synthesize accurate predictions of Revenue and Earnings Per Share (EPS) pertaining to various companies' stock releases.

Our experiments corroborate our findings from the previous chapter, in that our methods perform well in deriving Revenue and EPS predictions from Estimize data.

Similarly to the previous chapter, Section 5.1 of this chapter describes the dataset we use for our study. Section 5.2 describes our data features along with the methods we use. Section 5.3 presents our various experiments and associated analysis followed by a summary of our findings in Section 5.4.

5.1. Dataset

For our second study of prediction of events using Web information sources, we use data from Estimize, a platform that crowdsources earnings and economic estimates from independent and amateur analysts as well as those from hedge funds and brokerages. Specifically, the data consists of predictions of Revenue and EPS from various analysts for the stock releases *(releases)* of various companies.

	Training Dataset	Test Dataset
Analysts	2098	4185
Releases	2863	1222
Revenue Predictions	52244	40989
EPS Predictions	54648	59242
Total Predictions	106892	100231

Table 5.1.1. Estimize training and test datasets. The data ranges from January, 2012 to August 2016.

A platform such as Estimize can be subjected to manipulation by injection of predictions by users with malicious intent. As such, preprocessing of the data is performed in order to remove suspicious values: those within three Standard Deviations from the Mean. In addition, the platform commonly has repetitive entries by the same person for the same release, in which case, all but the latest is discarded.

The data also contains metadata pertaining to the analysts and the companies, such as analysts' biographical data and the industry a company belongs in, among others. However, such metadata is not used for our study. Additionally, the Estimize data contains Revenue and EPS predictions by Wall St.

The preprocessed Estimize data is then sorted in increasing chronological order, from the release with the earliest prediction to that with the latest. The timestamp used for a release is the one for the latest prediction for that release. The data is then partitioned into training and test partitions, ensuring that entire releases fall in either partition.

5.1.1. Grouping

As discussed later in Section 5.2.4, we use group sizes in range $N = [1, 13]$. Using our preprocessed Estimize data, we form datasets comprised of grouped predictions that we use for training and testing of our models. Each row of a dataset corresponds to a group formed by randomly batching $n \in N$ different predictions from the predictions available for the same release in the preprocessed data. For each n, we use Reservoir Sampling to identify up to m unique groups. Depending on n, the number of possible groups may be fewer than m, hence the relaxed requirement on how many groups are identified. The resulting datasets form our training dataset *(training set)* and test dataset *(test set)*. Table 5.1.1 shows details of the datasets.

The Estimize data is sourced courtesy of Dr. Ned Smith, a faculty member of the Kellogg School of Management at Northwestern University.

5.2. Methodology

Our task pertaining to Estimize is to develop a method to synthesize accurate financial predictions for releases from subsets of available predictions. Similarly to our task using Twitter data, this pursuit also tests our thesis that Web information can be harnessed to arrive at such predictions and also leverages the WoC to do so.

We begin the rest of this section by describing the models we apply to our data, followed by a discussion of our features.

5.2.1. Models

Similarly to our task with Twitter, we start by using majority vote (MV) to make predictions, and we use two variations of it. We also use machine learning (ML), for which we choose to use a model that is similar to Logistic Regression, but outputs continuous values as predictions as opposed to binary predictions as is required for game predictions with Twitter data. Therefore, our ML model of choice for the Estimize task is Linear Regression (LR).

5.2.2. Naïve Wisdom of Crowds

As we discuss shortly, our LR model is used with predictions by small groups of analysts. In order to compare our MV prediction performances models with those of our LR model, we test MV using two methods. One of the MV methods we use utilizes all available analysts to predict for a release. We call it *Naïve WoC*.

5.2.3. Selective Wisdom of Crowds

For our other method of MV, we use the same number of analysts as used by the LR model, i.e., MV predictions are made using groups of the same sizes as used by LR, each group formed by randomly selecting analysts. We call this method *Selective WoC*.

5.2.4. Linear Regression

To elaborate on the ML part of our task, our goal is to train a ML model on our data that learns to identify relatively small groups of analysts per release whose predictions can be aggregated to produce a prediction for that release. We identify such groups experimentally by training a LR model on a dataset comprising groups of analysts of sizes in range, $N = [1, 13]$, formed by batching analysts from the raw data. Specifically, the LR model learns to predict the magnitude of error in the prediction for a release by each group, upon which the groups are sorted by their error to identify those with the least predicted errors. Once a set of *candidate groups* (groups to be aggregated) is identified, we experiment again with differently sized sets or subgroups of the candidate groups to arrive at a final prediction, which is simply an average of the predictions by the candidate groups chosen to predict for the release.

Feature Name	Feature Description	Computed For
$Credibility_{grp}$	Group δ_g	Rev, EPS
$Contrarianness_{grp}$	Group Contrarianness	Rev, EPS
$PredictionDiversity_{grp}$	Diversity of Predictions	Rev, EPS, Combined
$AnalystBoPDiversity_{grp}$	Diversity of Analysts' BoP	Rev, EPS, Combined
$AnalystBoPSizeDiversity_{grp}$	Diversity of Analysts' BoP Sizes	Rev, EPS, Combined
$AnalystBoDPDiversity_{grp}$	Diversity of Analysts' BoPD	Rev, EPS
$PredictionHomogeneity_{grp}$	Homogeneity of Predictions	Rev, EPS
$GroupSize_{grp}$	Group Size	—
$CrowdPrediction_{grp}$	Average of Predictions by Group	Rev, EPS
$WallStPrediction_{grp}$	Prediction by Wall St.	Rev, EPS
$Actual_{grp}$	Actual Value	Rev, EPS
(Label) $NormalizedError_{grp}$	Group Normalized Error	Rev, EPS

Table 5.2.1. Estimize Features

5.2.5. Features

Our inquiry into synthesizing accurate financial predictions from subsets of crowd predictions is founded on the same hypotheses underlying our task using Twitter data, resulting in similar features measuring em credibility, contrarianness, and diversity. We also find the Estimize dataset to be particularly suitable to the application of WoC, in that the the dataset comprises numeric estimates from crowds of people, which the WoC phenomenon has been observed to perform well on. The viability of WoC motivates the use of groups of analysts for prediction, and the features we compute apply to the groups rather than to individual analysts. For example, when measuring credibility, we are interested mainly in the average credibility of a group of analysts.

We continue this section with a description of the features. Each feature is computed for both Revenue and EPS, the two prediction types. For certain features, a third instance of the feature is computed that combines Revenue and EPS information. The features are shown in Table 5.2.1.

5.2.5.1. Feature $Credibility_{grp}$. The credibility of a group of analysts is computed using the credibility measurement of every individual analyst in the group, $Credibility_a$, which is our metric for measuring an analyst's expertise by assessing their history of predictions. Unlike with the credibility measure used for the Twitter task, the credibility of an individual analyst is not agnostic of time, as the Estimize data spans a much longer period of time than does the Twitter data, allowing analysts to possibly improve in their predictions over time. As such, an analyst's history of predictions is limited to only their *past* predictions. $Credibility_a$ at a given time is the ratio of the cumulative prediction error made by an analyst over all their past predictions to the cumulative total of the actual prediction values. Subsequently, for every group of analysts, the group's expertise $Credibility_{grp}$ is computed as an average of the analysts' $Credibility_a$.

Specifically, for every analyst, $a \in A$ analysts with a prediction each for every release, $r \in R$ releases in the dataset prior to time, t, their $Credibility_a$ at time t is given by

$$Credibility_a = \frac{1}{|R|} \sum_{r \in R} \frac{|pred_{a_r} - true_r|}{|true_r| + 1e^{-6}} \tag{5.1}$$

where $pred_{a_r}$ denotes the analyst's prediction for the rth prior release, $true_r$ denotes the actual value of the same release, and the second component in the denominator prevents division by zero. Additionally, $Credibility_a$ is smoothed as follows using smoothing prior, m, and smoothing strength, λ:

$$Smoothed\,Credibility_a = \frac{Credibility_a|R| + \lambda m}{|R| + \lambda} \tag{5.2}$$

Therefore, for a group of predictions, grp, by A analysts, $Credibility_{grp}$ is given by

$$Credibility_{grp} = \frac{\sum\limits_{a \in A} Credibility_a}{|A|} \tag{5.3}$$

and is computed separately for both Revenue and EPS.

The **higher** the $Credibility_{grp}$ of a group of analysts is, the **less** cumulative error they have made in their past predictions of Revenue or EPS.

5.2.5.2. Feature $Contrarianness_{grp}$. Similarly to credibility measurement, the contrarianness of a group of analysts is computed using the contrarianness measurement of every individual analyst in the group, $Contrarianness_a$, based on their *past* predictions. $Contrarianness_a$, at a given time, is the degree to which an analyst has differed in the past from other analysts in predicting the same releases. At the time of an analyst's prediction for a release that is also predicted for by other analysts, the feature is the average of the differences between the analyst's predictions for prior releases and the other analysts' average predictions for the same releases. Subsequently, for every group of analysts, the group's contrarianness $Contrarianness_{grp}$ is computed as an average of the analysts' $Contrarianness_a$.

Specifically, for every analyst, $a \in A$ analysts with a prediction each for every release, $r \in R$ releases prior to time, t, each release with predictions by A' other analysts, where A' may be different for different r, their $Contrarianness_a$ at time, t is given by

$$Contrarianness_a = \frac{\sum\limits_{r \in R} |pred_{a_r} - \frac{1}{A'} \sum\limits_{a' \in A'} pred_{a'_r}|}{|R|} \tag{5.4}$$

where $pred_{a_r}$ denotes the analyst's prediction for the rth prior release, and $pred_{a'_r}$ denotes the prediction by every analyst, $a' \in A'$ for the same release. Additionally, $Contrarianness_a$ is smoothed as follows using smoothing prior, m, and smoothing strength, λ:

$$SmoothedContrarianness_a = \frac{Contrarianness_a|R| + \lambda m}{|R| + \lambda} \tag{5.5}$$

Therefore, for a group of predictions, grp, by A analysts, $Contrarianness_{grp}$ is given by

$$Contrarianness_{grp} = \frac{\sum\limits_{a \in A} Contrarianness_a}{|A|} \tag{5.6}$$

and is computed separately for both Revenue and EPS.

The **higher** the $Contrarianness_{grp}$ of a group of analysts is, the **more** contrarian they have been in their past predictions of Revenue or EPS.

5.2.5.3. Feature $PredictionDiversity_{grp}$. For a group of predictions by different analysts for a particular release, this feature measures the diversity of the predictions.

Specifically, for a group of predictions, grp, by A analysts, $PredictionDiversity_{grp}$ of the group is the Variance of the predictions and is given by

$$PredictionDiversity_{grp} = \frac{\sum\limits_{a \in A} (pred_a - \mu_{grp})^2}{|A|} \tag{5.7}$$

where $pred_a$ denotes the prediction by every analyst, $a \in A$, and μ_{grp} denotes the average of all predictions in grp. Additionally, $PredictionDiversity_{grp}$ is smoothed as follows using smoothing prior, m, and smoothing strength, λ:

$$SmoothedPredictionDiversity_{grp} = \frac{PredictionDiversity_{grp}|A| + \lambda m}{|A| + \lambda} \tag{5.8}$$

and is computed separately for Revenue and EPS.

The **higher** the $SmoothedPredictionDiversity_{grp}$ of a group of Revenue or EPS predictions is, the **more** variance there is in the predictions.

5.2.5.4. Feature $AnalystBoPDiversity_{grp}$.

For a group of predictions by different analysts for a particular release, this feature measures the diversity of the group of analysts based on the diversity of every individual analyst as compared to the other analysts in the group. It is computed in two steps using the Bag of Predictions (BoP) of every analyst, which is a binary vector that has a 1 for every release predicted by the analyst and a 0 otherwise. First, for every analyst in the group, their individual diversity is computed as the smoothed average of the Cosine Similarities between the BoP of the analyst and the BoP of every other analyst in the group. Next, $AnalystBoPDiversity_{grp}$ is computed as the smoothed average of the diversity measures of all the analysts in the group.

Specifically, for a group of predictions, grp, by A analysts, for every analyst, $a \in A$ with a BoP vector, v_a, their individual diversity $AnalystBoPDiversity_a$ is given by

$$AnalystBoPDiversity_a = \frac{\sum\limits_{a' \in A'} CosineSimilarity(v_a, v_{a'})}{|A'|} \tag{5.9}$$

where v_a denotes the BoP vector of every analyst $a \in A$, and $v_{a'}$ denotes the BoP vector of every analyst $a' \in A'$ other analysts in the group. Additionally, $AnalystBoPDiversity_a$ is smoothed as follows using smoothing prior, m, and smoothing strength, λ:

$$SmoothedAnalystBoPDiversity_a = \frac{AnalystBoPDiversity_a|A'| + \lambda m}{|A'| + \lambda} \tag{5.10}$$

Therefore, $AnalystBoPDiversity_{grp}$ of the group is given by

$$AnalystBoPDiversity_{grp} = \frac{\sum\limits_{a \in A} SmoothedAnalystBoPDiversity_a}{|A|} \tag{5.11}$$

Additionally, $AnalystBoPDiversity_{grp}$ is smoothed as follows using smoothing prior, m, and smoothing strength, λ:

$$SmoothedAnalystBoPDiversity_{grp} = \frac{AnalystBoPDiversity_{grp}|A| + \lambda m}{|A| + \lambda} \tag{5.12}$$

and is computed for Revenue, EPS, and a combination of both, for which the analyst vectors used have 1 for releases for which the analyst has predicted *both* Revenue and EPS and 0 otherwise.

The **higher** the $SmoothedAnalystBoPDiversity_{grp}$ of a group of Revenue or EPS predictions is, the **more** diverse the analysts who made those predictions have been in their past common predictions.

5.2.5.5. Feature $AnalystBoPSizeDiversity_{grp}$. For a group of predictions by different analysts for a particular release, this feature, similarly to $AnalystBoPDiversity_{grp}$, measures the diversity of the analysts based on the diversity of every individual analyst as compared to the other analysts in the group. However, unlike with $AnalystBoPDiversity_{grp}$, the diversity of every individual analyst is not computed using their BoP, but the *size* of their BoP. As such, the computation of the feature is more straightforward and is simply the Variance of the BoP sizes of the analysts.

Specifically, for a group of predictions, grp, by A analysts, with every analyst, $a \in A$ with a BoP vector, v_a, $AnalystBoPSizeDiversity_{grp}$ is given by

$$AnalystBoPSizeDiversity_{grp} = \frac{\sum\limits_{a \in A} (|v_a| - \mu_{|v_{a \in A}|})^2}{|A|} \tag{5.13}$$

where $|v_a|$ is the BoP size of every analyst, $a \in A$, and $\mu_{|v_{a \in A}|}$ is the average BoP size of all the analysts in the group. Additionally, $AnalystBoPSizeDiversity_{grp}$ is smoothed as follows using smoothing prior, m, and smoothing strength, λ:

$$Smoothed\, AnalystBoPSizeDiversity_{grp} = \frac{AnalystBoPSizeDiversity_{grp}|A| + \lambda m}{|A| + \lambda} \tag{5.14}$$

and is computed separately for Revenue, EPS, and a combination of both, for which the analyst vectors used have 1 for releases for which the analyst has predicted *both* Revenue and EPS and 0 otherwise.

The **higher** the $Smoothed\,AnalystBoPSizeDiversity_{grp}$ of a group of Revenue or EPS predictions is, the **more** diverse the analysts who made those predictions have been in their past common predictions.

5.2.5.6. Feature $AnalystBoDPDiversity_{grp}$. For a group of predictions by different analysts for a particular release, this feature, similarly to $AnalystBoPDiversity_{grp}$, measures the diversity of the analysts based on the diversity of every individual analyst as compared to the other analysts in the group, and is computed in the same two-step method. However, instead of using the BoP of every analyst, it uses the *Bag of Directional Predictions (BoDP)* of every analyst. BoDP differs from BoP in that it is ternary as opposed to binary, using 1 to indicate releases for which predictions are equal to or higher than the prediction by Wall Street, –1 to indicate releases for which predictions are lower than that by Wall Street, and 0 to indicate releases for which there are no predictions. First, for every analyst in the group, their individual diversity is computed as the smoothed average of the Cosine Similarities between the BoDP of the analyst and the BoDP of every other analyst in the group. Next, $AnalystBoDPDiversity_{grp}$ is computed as the smoothed average of the diversity measures of all the analysts in the group.

Specifically, for a group of predictions, grp, by A analysts, for every analyst, $a \in A$ with a BoDP vector, v_a, their individual diversity $AnalystBoDPDiversity_a$ is given by

$$AnalystBoDPDiversity_a = \frac{\sum\limits_{a' \in A'} CosineSimilarity(v_a, v_{a'})}{|A'|} \tag{5.15}$$

where v_a denotes the BoDP vector of every analyst, $a \in A$, and $v_{a'}$ denotes the BoDP vector of every analyst, $a' \in A'$ other analysts in the group. Additionally, $AnalystBoDPDiversity_a$ is smoothed as follows using smoothing prior, m, and smoothing strength, λ:

$$Smoothed\,AnalystBoDPDiversity_a = \frac{AnalystBoDPDiversity_a|A'| + \lambda m}{|A'| + \lambda} \tag{5.16}$$

Therefore, $AnalystBoDPDiversity_{grp}$ of the group is given by

$$AnalystBoDPDiversity_{grp} = \frac{\sum\limits_{a \in A} Smoothed\,AnalystBoDPDiversity_a}{|A|} \tag{5.17}$$

Additionally, $AnalystBoDPDiversity_{grp}$ is smoothed as follows using smoothing prior, m, and smoothing strength, λ:

$$Smoothed\,AnalystBoDPDiversity_{grp} = \frac{AnalystBoDPDiversity_{grp}|A| + \lambda m}{|A| + \lambda} \tag{5.18}$$

and is computed separately for both Revenue and EPS.

The **higher** the $Smoothed\,AnalystBoDPDiversity_{grp}$ of a group of Revenue or EPS predictions is, the **more** diverse the analysts who made those predictions have been directionally in their past common predictions.

5.2.5.7. Feature $PredictionHomogeneity_{grp}$. For a group of predictions by different analysts for a particular release, this feature measures the homogeneity of the predictions with respect to their direction in comparison with the prediction by Wall St.

Specifically, for a group of predictions, grp, by A analysts, the $PredictionHomogeneity_{grp}$ of the group is the ratio of the maximum number of predictions that are directionally equivalent to the total number of predictions. Let $count(x)$ be the number of predictions in grp that are equal to x. Then, $PredictionHomogeneity_{grp}$ is given by

$$PredictionHomogeneity_{grp} = \frac{\max\{count(1), count(0), count(-1)\}}{|A|} \tag{5.19}$$

where 1, 0, and –1 denote, respectively, a prediction higher than that by Wall St., a prediction equal to that by Wall St., and a prediction lower than that by Wall St. Additionally, $PredictionHomogeneity_{grp}$ is smoothed as follows using smoothing prior, m, and smoothing strength, λ:

$$Smoothed\,PredictionHomogeneity_{grp} = \frac{PredictionHomogeneity_{grp}|A| + \lambda m}{|A| + \lambda} \tag{5.20}$$

and is computed separately for both Revenue and EPS.

The **higher** the $Smoothed\,PredictionHomogeneity_{grp}$ of a group of Revenue or EPS predictions is, the **more** homogeneous the predictions are.

5.2.5.8. Feature $GroupSize_{grp}$. For a group of predictions by different analysts for a particular release, this feature is simply the size of the group of predictions.

Specifically, for a group of predictions, grp, by A analysts, the $GroupSize_{grp}$ of the group is given by

$$GroupSize_{grp} = |group| \tag{5.21}$$

Three features of the data inform a set of *Prediction Features*: For a particular release, they are predictions made by the analysts, predictions made by Wall St., and the actual value of Revenue or EPS of the release. These features are discussed below.

5.2.5.9. Feature $CrowdPrediction_{grp}$. For a group of predictions by different analysts for a particular release, this feature is the *wisdom of the crowd* and measures the prediction by the crowd as the average of the predictions by the analysts.

Specifically, for a group of predictions, grp, by A analysts, the $CrowdPrediction_{grp}$ of the group is given by

$$CrowdPrediction_{grp} = \frac{\sum\limits_{a \in A} pred_a}{|A|} \tag{5.22}$$

where $pred_a$ denotes the prediction by every analyst, $a \in A$.

$CrowdPrediction_{grp}$ is computed separately for both Revenue and EPS.

5.2.5.10. Feature $WallStPrediction_{grp}$. For a group of predictions by different analysts for a particular release, this feature is simply the prediction by Wall St for the release.

Specifically, for a group of predictions, grp, by A analysts, the $WallStPrediction_{grp}$ of the group is given by

$$WallStPrediction_{grp} = pred_{WallSt} \tag{5.23}$$

where $pred_{WallSt}$ denotes the prediction by Wall St. for the release.

$WallStPrediction_{grp}$ is computed separately for both Revenue and EPS.

5.2.5.11. Feature $Actual_{grp}$. For a group of predictions by different analysts for a particular release, this feature is simply the value of the Revenue or the EPS of the release.

Specifically, for a group of predictions, grp, by A analysts, the $Actual_{grp}$ of the group is given by

$$Actual_{grp} = actual \tag{5.24}$$

where $actual$ denotes the actual value of the Revenue or the EPS of the release.

$Actual_{grp}$ is computed separately for both Revenue and EPS.

5.2.5.12. Feature $NormalizedError_{grp}$. For a group of predictions by different analysts for a particular release, this feature is the *target* or *label* for the ML model to learn to predict and is a normalized measurement of error. The error measured is the delta between the crowd prediction and the actual value for the release, which is then normalized by dividing it by the actual value.

Specifically, for a group of predictions, grp, by A analysts, the $NormalizedError_{grp}$ of the group is given by

Feature Suites
All features (in Table 5.2.1)
Only non-Revenue features
Only non-EPS features
Only non-Credibility features
Only non-Contrarianness features
Only non-Diversity features
Only non-Homogeneity features

Table 5.2.2. Ablation Study Feature Suites

$$NormalizedError_{grp} = \frac{\frac{\sum_{a \in A} pred_a}{|A|} - actual}{actual} \tag{5.25}$$

Using Equations 5.22 and 5.24, the $NormalizedError_{grp}$ can be rewritten as

$$NormalizedError_{grp} = \frac{CrowdPrediction_{grp} - actual}{actual} \tag{5.26}$$

5.2.6. Ablation Study

The purpose of an ablation study is to assess the importance of individual features by contrasting the performances of a trained model when it is trained and tested on data that includes the feature and on data that excludes the feature. A variant of the method is to leave out an entire set of feature at a time if certain features can be grouped into a common category based on some aspect of the features. We perform several rounds of model training, each round performing an ablation study by leaving one feature out a time from a suite of features. Table 5.2.2 lists the suites of features used. Additionally, each round of ablation study is performed using several sizes for how many candidate groups with the lowest predicted error is to be used to aggregate their predictions into a final prediction.

5.2.7. Error Measurements

To gauge the performance of our trained model, we measure *accuracy*, which we define in terms of the error measurements we discuss below. Each of then provides a measure of accuracy of the prediction by a group.

5.2.7.1. Directional Accuracy. For the task of synthesizing accurate financial predictions from the wisdom of crowds, we consider predictions by Wall St. a good benchmark to compare to, as Wall St. is renowned for their robustly accurate financial predictions. As such, if a group's prediction is on the same side (higher or lower) of the actual value, we consider the prediction to have an important attribute of accuracy: directional correctness. We define *Directional Accuracy* ("*DirAcc*") for a group's prediction to be a *count* in [0, 1] such that it is 1 when the prediction is directionally correct and 0 otherwise. The value has two convenient properties: it serves as a binary indicator of directional accuracy and can also be summed across releases for a cumulative measure of directional accuracy. For a prediction, $pred_{grp}$, by a group for a release, r, if the actual value for the release is *actual* and the Wall St. prediction is $pred_{WallSt}$, DirAcc is given by

$$DirAcc = \begin{cases} 1, & \text{if } (pred_{grp} - actual)(pred_{WallSt} - actual) >= 0 \\ 0, & \text{otherwise} \end{cases} \tag{5.27}$$

5.2.7.2. Absolute Error. *Absolute Error (AE)* is the absolute value of the delta between a group's prediction and the actual value for a release. For a prediction, $pred_{grp}$, by a group for a release, r, the actual value for which is *actual*, AE is given by

$$AE = |pred_{grp} - actual| \tag{5.28}$$

5.2.7.3. 1% Error. We define *1% Error* ($1\%Error$) for a group's prediction to be a *count* in [0, 1] such that it is 1 when the prediction is within 1% of the actual value—the absolute value of the delta between a group's prediction and the actual value for a release is less than or equal to 1% of the actual value—and 0 otherwise. For a prediction, $pred_{grp}$, by a group for a release, r, with an actual value, $actual$, $1\%Error$ is given by

$$1\%Error = \begin{cases} 1, & \text{if } pred_{grp} - actual <= actual \times 1\% \\ 0, & \text{otherwise} \end{cases} \tag{5.29}$$

5.2.7.4. 5% Error. We define *5% Error* ($5\%Error$) for a group's prediction to be a *count* in [0, 1] such that it is 1 when the prediction is within 5% of the actual value—the absolute value of the delta between a group's prediction and the actual value for a release is less than or equal to 5% of the actual value—and 0 otherwise. For a prediction, $pred_{grp}$, by a group for a release, r, with an actual value, $actual$, $5\%Error$ is given by

$$5\%Error = \begin{cases} 1, & \text{if } pred_{grp} - actual <= actual \times 5\% \\ 0, & \text{otherwise} \end{cases} \tag{5.30}$$

5.2.7.5. 10% Error. We define *10% Error* ($10\%Error$) for a group's prediction to be a *count* in [0, 1] such that it is 1 when the prediction is within 10% of the actual value—the absolute value of the delta between a group's prediction and the actual value for a release is less than or equal to 10% of the actual value—and 0 otherwise. For a prediction, $pred_{grp}$, by a group for a release, r, with an actual value, $actual$, $10\%Error$ is given by

$$10\%Error = \begin{cases} 1, & \text{if } pred_{grp} - actual <= actual \times 10\% \\ 0, & \text{otherwise} \end{cases} \tag{5.31}$$

5.3. Experiments and Analysis

Based on our ablation studies, we test the performance of predictions for Estimize using a subset of all the features listed in Table 5.2.1. The features we use are those that show to yield the best performances in the ablation study. They are further informed by the correlation coefficients resulting from LR performed with all features. Also, if our selected features include Wall St. predictions, we repeat our experiments with those features omitted in order to study the influence of Wall St. predictions on our models.

Additionally, we compare the results with those achieved by using all features of the data.

We begin by looking at results from experiments performed for Revenue predictions.

5.3.1. Revenue Prediction

5.3.1.1. Selected Features. Table 5.3.1 shows the prediction performances when all 13 groups sizes are used.

We also want to study the effect of group size on the predictions, which is the motivation behind the use of the different group sizes. In preliminary experiments, we observe prediction

Model	Directional Accuracy	1% Error	5% Error	10% Error	Beats Wall St. in AE
Naive WoC	0.84	0.31	0.82	0.96	–
Selective WoC	**0.93**	0.34	**0.83**	**0.97**	**0.57**
LR	0.91	**0.35**	**0.83**	**0.97**	**0.57**
LR (Sans Wall St.)	**0.93**	0.34	0.82	**0.97**	**0.57**
Wall St.	–	0.32	0.82	0.96	–

Table 5.3.1. Estimize Revenue prediction performances with the best features: All group sizes. Performances are shown as ratio of average number of correct predictions to average number of total predictions (average of 485.38) across 13 different group sizes in range N = [1, 13]. Features used are $\text{Credibility}_{\text{grp}}$ EPS, $\text{Credibility}_{\text{grp}}$ Revenue, $\text{AnalystBoDPDiversity}_{\text{grp}}$ Revenue, $\text{WallStPrediction}_{\text{grp}}$ EPS, and $\text{WallStPrediction}_{\text{grp}}$ Revenue.

performances using groups sizes in the range eight to 12 to stand out with small amounts of data. Incidentally, Kao and Couzin observe the same effect in their research [**37**], about which one of the authors comment in a subsequent interview, “there’s a small optimal group size of eight to 12 individuals that tends to optimize decisions.” However, we do not observe the same effect in our Revenue predictions.

Both WoC and LR perform well overall, notably in Directional Accuracy—the prediction is considered accurate if it is on the same side of the Wall St. prediction as the actual release value. LR appears to outperform the other predictions, albeit by a narrow margin. While LR outperforms the other models in all the error categories, we note the presence of 1% Error among them, as it represents the least error. We also note that LR trades that level of performance off with Directional Accuracy. It is not apparent, though, if such tradeoff is an artifact of noise or have significance.

Looking closely at the results in Table 5.3.1, we first note the features that achieved the LR results. We observe that the results—the best among all our trials with different subsets of features—require only a few features and that they pertain mostly to *credibility* and Wall St. predictions. This does not surprise us, as relying on the credibility of the predicting crowd is one of the core tenets of our thesis. The role of the Wall St. predictions is not apparent, but we presume it serves as an anchor for the LR model to learn, guiding its prediction when the Wall St. prediction is taken into account in conjunction with the crowd predictions.

We also notice the presence of a *diversity* feature in the subset of features in discussion. The specific diversity feature here is $\text{AnalystBoDPDiversity}_{\text{grp}}$, which measures the diversity of the directional predictions of the analysts in a group. As diversity is one of the pillars of the WoC, as put forth by Surowiecki in [**23**], this feature also seems perfectly reasonable to be included in the feature subset. We are curious to know if the directionality aspect of the predictions that this feature pertains to is important in that directional diversity results in the canceling effect of predictions that is thought to be behind the WoC phenomenon. We slate the inquiry as future work.

Selective WoC, which uses the same number of resources (analysts) used by LR, outperforms Naive WoC, and by a large margin in Directional Accuracy. Such effects are interesting and warrant further study in future research.

One surprising observation is that Wall St. predictions do not outperform WoC or LR on any metric, though we recognize its performance difference with the other models to be small.

Lastly, we look at the last column in the table, which captures the ratio of predictions for which Selective WoC and LR outperform Wall St. in Absolute Error (AE). The 1% through 10% Error metrics provide some margin for accuracy and may favor one metric or another depending on how close to and which side of a thresholds the accuracy is. AE gives a different aspect of accuracy in that it capture the absolute cumulative delta between model

Model	Directional Accuracy	1% Error	5% Error	10% Error	Beats Wall St. in AE
Naive WoC	0.84	0.31	0.82	0.96	–
Selective WoC	**0.93**	**0.34**	**0.83**	**0.97**	**0.57**
LR	0.90	0.32	0.82	0.96	**0.57**
LR (Sans Wall St.)	0.90	0.33	**0.83**	0.96	**0.57**
Wall St.	–	0.32	0.82	0.96	–

Table 5.3.2. Estimize Revenue prediction performances with all features: All group sizes. Performances are shown as ratio of average number of correct predictions to average number of total predictions (average of 485.38) across 13 different group sizes in range $N = [1, 13]$. All features are used.

predictions and actual release values. We note that all three models represented in the last column has close to 60% accuracy as measured by AE.

Overall, the performances may be interpreted to lend credibility to the notion that it is plausible to identify data features that allow aggregation of a subset of crowd predictions to make predictions for domains such as that pertaining to Estimize. Further, these results show that WoC, if applied selectively, may be sufficient for the prediction task, lending further credibility to the WoC phenomenon.

5.3.1.2. All Features. The results shown thus far are based on features identified in the ablation study as those that perform the best. We now compare the results with those achieved by using all features of the data as listed in Table 5.2.1.

Table 5.3.2 shows Estimize prediction performances when all groups are used.

The lack of selectiveness in the features used seems to have an adverse effect on the LR predictions, though the differences are small. Considering that the loss in accuracy is seen across all metrics, it seems reasonable to conclude that identifying a subset of optimal features to use for prediction is preferred to using all available features.

We perform some additional experiments to study certain features in isolation by assessing their impact on performance. We find that, contrary to our observations in preliminary experiments pertaining to the Twitter task, *contrarianness* does not seem to have an effect on prediction accuracy. The same can be stated for diversity features other than the one discussed in this section and for *homogeneity*.

5.3.2. EPS Prediction

5.3.2.1. Selected Features. Table 5.3.3 shows the prediction performances when all 13 groups sizes are used.

WoC and LR both perform slightly better than Wall St. predictions, but do not exhibit strong performance overall except in Directional Accuracy. We immediately notice that, unlike LR performance for Revenue prediction, LR appears to do better, though only slightly, when Wall St. predictions are omitted as opposed to when they are included. Once again, such minor differences may not be significant, but the effect is observed across all metrics.

More noteworthy is the set of features for which LR yields the accuracies, which are the best from several experiments that test different sets of feature combinations. We note that *contrarianness* features are present in the subset of features that yield the accuracies in discussion. This reinvigorates the question of whether contrarians are better predictors. Also, *credibility* features are part of the feature subset as well, as was the case with Revenue predictions. It appears credibility may be a strong indicator of prediction accuracy. Further investigation into its influence may reveal more insights.

We note once again that Wall St. lags behind other models, though only by a small margin.

Model	Directional Accuracy	1% Error	5% Error	10% Error	Beats Wall St. in AE
Naïve WoC	0.69	0.09	0.38	0.61	–
Selective WoC	**0.74**	0.10	0.39	**0.62**	**0.57**
LR	0.72	0.10	0.39	0.61	**0.57**
LR (Sans Wall St.)	0.73	**0.11**	**0.40**	0.61	**0.57**
Wall St.	–	0.10	0.37	0.60	–

Table 5.3.3. Estimize EPS prediction performances with the best features: All group sizes. Performances are shown as ratio of average number of correct predictions to average number of total predictions (average of 654.62) across 13 different group sizes in range $N = [1, 13]$. Features used are Credibility$_{\text{grp}}$ EPS, Credibility$_{\text{grp}}$ Revenue, Contrarianness$_{\text{grp}}$ EPS, and Contrarianness$_{\text{grp}}$ Revenue.

Model	Directional Accuracy	1% Error	5% Error	10% Error	Beats Wall St. in AE
Naïve WoC	0.69	0.09	0.38	0.61	–
Selective WoC	**0.74**	**0.10**	**0.40**	**0.62**	**0.57**
LR	0.73	0.**10**	0.39	0.61	**0.57**
LR (Sans Wall St.)	**0.74**	**0.10**	0.39	0.61	**0.57**
Wall St.	–	**0.10**	0.37	0.60	–

Table 5.3.4. Estimize EPS prediction performances with all features: All group sizes. Performances are shown as ratio of average number of correct predictions to average number of total predictions (average of 654.6) across 13 different group sizes in range $N = [1, 13]$. All features are used.

5.3.2.2. All Features. Table 5.3.4 shows EPS prediction performances when all features from Table 5.2.1 are used.

When the performances are evaluated using the group size aggregation categories mentioned earlier, we do not observe any noteworthy effect of the aggregations in our EPS predictions, as is also the case with Revenue predictions.

Finally, additional experiments to study certain features in isolation by assessing their impact on performance do not show any other features to have significant influence on prediction accuracy.

5.4. Summary of Findings

In summary, the findings from our research on synthesizing financial predictions are as follow:

(1) LR and MV consistently outperform Wall St. across all metrics and for both Revenue and EPS, even though the performance differences are small. Given a benchmark such as Wall St., we conclude our methods certainly demonstrate viability and lend support to our thesis that people's predictions can be aggregated in order to synthesize all their voices into a singular predictive voice.

(2) The fraction of times LR and WoC beat Wall St. in cumulative absolute error is encouraging and lends further support to our thesis and methods.

(3) Credibility, contrarianness, and diversity all stand out as important feature types. We believe further study of the pertinent features will prove very insightful.

(4) The high performance of our models in directional prediction is noteworthy, as being closer to the actual value than Wall St. is not a trivial feat.

CHAPTER 6

Future Work and Conclusion

In this dissertation, we present research that we conducted to test our thesis that Web information sources can be harnessed to predict event outcomes. Our work is motivated by the need to sift information from misinformation in order to arrive at the truth when faced with noisy Web information that may lead to misbeliefs and wrong prognostications. We adopt an approach to solve the problem that acknowledges the presence of misinformation amidst information and investigates methods to derive at accurate predictions of events in order to arrive at the truth.

As part of our two-fold approach, we first measure a particular bias: the content emphasis of Web search engines, Google and Bing. We define content emphasis as the degree to which differences across search engines' rankings correlate with features of the ranked content. We find no evidence of emphasis in the search engines' rankings that express positive orientation toward the engine company's products. We do find, however, that Google slightly emphasizes negative results in general, on our dataset. We also find that the engines emphasize particular news sites over others, with Google showing a tendency to favor smaller news outlets while Bing favoring bigger ones. Compared to Google, Bing also favors content with Facebook Like buttons. Furthermore, both engines favor pages containing their own vendor company's advertisements as opposed to competitors' advertisements. Along with the aforementioned findings, our contributions from this research include a first-of-its-kind system, PAWS, a Platform for Analyzing Web Search engine.

For the second part of our research, we rely on predictions crowdsourced from the Web to synthesize accurate predictions of event outcomes. Specifically, we use game predictions from Twitter pertaining to four major sporting tournaments and financial predictions from prediction market, Estimize, about Revenue and Earnings Per Share related to stock releases of various companies. We rely on the wisdom of crowds (WoC), the phenomenon that the average of various estimates pertaining to a topic by a crowd of people tends to be more accurate than the estimate by any one individual in the crowd, often matching or even exceeding the accuracy of estimates by experts. Using majority vote as our WoC method and Logistic Regression as a machine learning approach, we show that it is possible, in certain cases, to arrive at accurate predictions of game outcomes. We find evidence of the important of certain features of data that aid in prediction. However, the relationship among the features are non-trivial and warrant further study to gain more insights on their nuanced behavior, which should lead to more accurate predictions. With the Estimize data, we use majority vote and Linear Regression to synthesize predictions and compare them with predictions by Wall St., which we consider expert predictions for our scenario. We show that both majority vote and Linear Regression outperform Wall St. when used with small groups or subsets of predictions from among a much larger corpus.

We recognize the need for further studies to investigate the intricate relationships and characteristics of our data features and the data corpuses themselves to derive more accurate predictions at scale. We slate such inquiries for future research to continue to better understand WoC and event prediction toward the goal of uncovering truth in Web information. Our research leads to a number of questions that can form the basis for fascinating future work. From our very earliest experiments, the contrarianness of individuals whose predictions we use has indicated the possibility of its having a role in predictions. While we have not found strong evidence of contrarianness' being important, we have not found evidence of the contrary either. Does contrarianness drive prediction accuracy? Diversity is considered a core tenet of the WoC phenomenon. What are ways to measure diversity outside of the various methods we use in our research? Does the size of the group used for prediction seem to play a role in predictions as observed in other research? These are questions that we find very motivated to answer, and we are very encouraged by our own findings to follow through with future research for the very goal of answering them.

CHAPTER 7

Appendix

Tables begin on the next page.

Game	All Games		Game	Top Games	
	LR	MV		LR	MV
sri lanka-new zealand	0.72	0.69	england-australia	0.50	0.55
england-australia	0.45	0.53	india-pakistan	0.52	0.32
india-pakistan	0.78	0.38	new zealand-scotland	1	0.91
south africa-zimbabwe	0.92	0.87	bangladesh-afghanistan	0	0.09
west indies-ireland	0.82	0.90	australia-bangladesh	0.33	0
new zealand-scotland	0.95	0.95	pakistan-west indies	0.08	0.19
bangladesh-afghanistan	0.02	0.07	south africa-india	0.85	0.77
zimbabwe-united arab emirates	0.63	0.52	south africa-west indies	0.90	0.87
australia-bangladesh	0.27	0.50	pakistan-zimbabwe	0.97	0.75
new zealand-england	0.55	0.50	south africa-ireland	0.62	0.40
sri lanka-afghanistan	0.20	0.28	pakistan-united arab emirates	0.97	0.88
pakistan-west indies	0.15	0.18	new zealand-afghanistan	0.92	0.89
south africa-india	0.93	0.79	australia-sri lanka	0.88	0.79
england-scotland	0.40	0.52	zimbabwe-ireland	0.15	0.28
west indies-zimbabwe	0.63	0.34	england-bangladesh	0.52	0.64
ireland-united arab emirates	0.42	0.26	south africa-pakistan	0.97	0.90
sri lanka-bangladesh	0.85	0.70	india-ireland	0.92	0.86
australia-new zealand	0.69	0.46	south africa-united arab emirates	0.92	0.80
south africa-west indies	0.92	0.84	india-zimbabwe	0.98	0.91
afghanistan-scotland	0.52	0.40	australia-scotland	0.98	0.92
pakistan-zimbabwe	0.93	0.81	pakistan-ireland	0.93	0.86
england-sri lanka	0.48	0.63	bangladesh-new zealand	0.53	0.58
south africa-ireland	0.58	0.41	england-afghanistan	0.88	0.82
bangladesh-scotland	1	0.98	sri lanka-south africa	0.43	0.54
pakistan-united arab emirates	0.92	0.84	bangladesh-india	0.75	0.73
new zealand-afghanistan	0.97	0.90	australia-pakistan	0.02	0.22
australia-sri lanka	0.95	0.78	new zealand-west indies	0.97	0.86
zimbabwe-ireland	0.23	0.24	new zealand-south africa	0.50	0.43
england-bangladesh	0.58	0.58	australia-india	0.25	0.27
south africa-pakistan	0.92	0.87	new zealand-australia	0.57	0.40
india-ireland	0.93	0.80			
south africa-united arab emirates	0.92	0.78			
sri lanka-scotland	0.59	0.63			
india-zimbabwe	0.97	0.87			
australia-scotland	0.97	0.92			
pakistan-ireland	0.97	0.85			
bangladesh-new zealand	0.52	0.46			
west indies-united arab emirates	0.54	0.67			
england-afghanistan	0.88	0.82			
sri lanka-south africa	0.63	0.52			
bangladesh-india	0.88	0.79			
australia-pakistan	0.05	0.20			
new zealand-west indies	0.95	0.86			
new zealand-south africa	0.57	0.45			
australia-india	0.43	0.35			
new zealand-australia	0.67	0.50			
Average	0.67	0.61	Average	0.66	0.61

Table 7.0.1. Cricket Logistic Regression (LR) Results. LR accuracies vs. MV accuracies, for all 46 games on the left (*All Games*) and the top 30 games on the right (*Top Games:* those with 90 or more predictions).

Game	All Games		Game	Top Games	
	LR	MV		LR	MV
spain-netherlands	0.73	0.28	spain-netherlands	0.82	0.22
belgium-russia	0.70	0.32	belgium-russia	0.77	0.25
usa-germany	0.67	0.37	usa-germany	0.75	0.36
ghana-usa	0.73	0.41	nigeria-argentina	0.67	0.37
italy-uruguay	0.79	0.42	colombia-ivory coast	0.65	0.42
argentina-iran	0.53	0.42	ghana-usa	0.77	0.44
colombia-ivory coast	0.65	0.43	italy-costa rica	0.63	0.44
nigeria-argentina	0.58	0.45	uruguay-england	0.52	0.46
italy-costa rica	0.58	0.45	argentina-iran	0.58	0.50
nigeria-bosnia and herzegovina	0.53	0.51	australia-spain	0.50	0.52
cameroon-brazil	0.47	0.52	bosnia and herzegovina-iran	0.50	0.52
australia-spain	0.52	0.53	cameroon-brazil	0.53	0.52
uruguay-england	0.58	0.53	nigeria-bosnia and herzegovina	0.52	0.54
bosnia and herzegovina-iran	0.53	0.53	honduras-ecuador	0.47	0.55
croatia-mexico	0.45	0.56	croatia-mexico	0.47	0.57
honduras-ecuador	0.52	0.59	netherlands-chile	0.32	0.57
netherlands-chile	0.43	0.60	korea-algeria	0.33	0.58
korea-belgium	0.42	0.64	germany-portugal	0.35	0.61
korea-algeria	0.58	0.64	switzerland-france	0.08	0.75
greece-ivory coast	0.56	0.69	portugal-ghana	0.10	0.76
germany-portugal	0.32	0.71	belgium-algeria	0.05	0.89
honduras-switzerland	0.67	0.72			
switzerland-france	0.67	0.78			
japan-colombia	0.88	0.82			
portugal-ghana	0.70	0.83			
belgium-algeria	0.65	0.85			
mexico-cameroon	0.90	0.92			
Average	0.61	0.57	Average	0.49	0.52

Table 7.0.2. Soccer Logistic Regression (LR) Results. LR accuracies vs. MV accuracies, for all 46 games on the left (*All Games*) and the top 30 games on the right (*Top Games:* those with 90 or more predictions).

Game	All Games		Game	Top Games	
	LR	MV		LR	MV
vikings-49ers	1	0	lions-cowboys	1	0.08
titans-browns	1	0	ravens-patriots	1	0.09
rams-vikings	1	0	seahawks-packers	1	0.05
dolphins-bills	1	0	packers-cowboys	0.98	0.22
dolphins-jets	1	0	cardinals-panthers	0.98	0.01
lions-rams	1	0	bengals-colts	0.85	0.26
falcons-jaguars	1	0	seahawks-patriots	0.65	0.5
cardinals-panthers	1	0.02	ohio state-alabama	0.37	0.51
lions-cowboys	1	0.05	florida state-oregon	0.33	0.52
seahawks-rams	1	0.15			
titans-saints	1	0.23			
cardinals-seahawks	1	0.37			
bears-lions	1	0.46			
chiefs-vikings	1	0.5			
vikings-lions	1	0.67			
texans-jaguars	1	0.83			
vikings-raiders	1	0.83			
bengals-bills	1	1			
buccaneers-falcons	1	1			
rams-bengals	1	1			
...	...	...			
texans-bengals	0.14	0.73			
chargers-chiefs	0.14	0.79			
buccaneers-redskins	0.14	0.83			
texans-colts	0.13	0.69			
chiefs-texans	0.11	0.51			
bills-chiefs	0	0			
titans-jets	0	0			
raiders-titans	0	0.17			
cardinals-49ers	0	0.17			
chiefs-broncos	0	0.5			
vikings-falcons	0	0.5			
dolphins-chargers	0	0.53			
giants-buccaneers	0	0.67			
jets-texans	0	0.81			
vikings-bears	0	0.83			
saints-redskins	0	0.96			
raiders-chargers	0	1			
bears-rams	0	1			
49ers-seahawks	0	1			
bengals-49ers	0	1			
Average	0.54	0.5	Average	0.8	0.25

Table 7.0.3. Football Logistic Regression (LR) Results. LR accuracies vs. MV accuracies, for all 170 games on the left (*All Games*) (only the top 20 and the bottom 20 as sorted by LR accuracy shown due to limited space) and the top 9 games on the right (*Top Games:* those with 90 or more predictions).

Bibliography

[1] Xiaoxin Yin, Jiawei Han, and S Yu Philip. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808, 2008.

[2] Matthew Richardson, Rakesh Agrawal, and Pedro Domingos. Trust management for the semantic web. In *International semantic Web conference*, pages 351–368. Springer, 2003.

[3] Jennifer Golbeck, Bijan Parsia, and James Hendler. Trust networks on the semantic web. In *International Workshop on Cooperative Information Agents*, pages 238–249. Springer, 2003.

[4] Min Jiang. Search concentration, bias, and parochialism: A comparative study of google, baidu, and jike's search results from china. *Journal of Communication*, 64(6):1088–1110, 2014.

[5] Ryen W White and Ahmed Hassan. Content bias in online health search. *ACM Transactions on the Web (TWEB)*, 8(4):25, 2014.

[6] Alexandre Cornière and Greg Taylor. Integration and search engine bias. *The RAND Journal of Economics*, 45(3):576–597, 2014.

[7] Juwon Kwak. Has the dominant search engine, if any, discriminated against rival websites? *If Any, Discriminated Against Rival Websites*, 2015.

[8] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.

[9] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*, pages 228–243. Springer, 2014.

[10] Sujoy Kumar Sikdar, Byungkyu Kang, John O?Donovan, Tobias Hollerer, and Sibel Adal. Cutting through the noise: Defining ground truth in information credibility on twitter. *HUMAN*, 2(3):151–167, 2013.

[11] Mohammed A Alam and Doug Downey. Analyzing the content emphasis of web search engines. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 1083–1086. ACM, 2014.

[12] Sergiu Chelaru, Ismail Sengor Altingovde, Stefan Siersdorfer, and Wolfgang Nejdl. Analyzing, detecting, and exploiting sentiment in web queries. *ACM Trans. Web*, 8(1):6:1–6:28, December 2013.

[13] Gianluca Demartini and Stefan Siersdorfer. Dear search engine: What's your opinion about...?: Sentiment analysis for semantic enrichment of web search results. In *Proceedings of the 3rd Int'l Semantic Search Workshop*, SEMSEARCH '10, pages 4:1–4:7, New York, NY, USA, 2010. ACM.

[14] Abbe Mowshowitz and Akira Kawaguchi. Measuring search engine bias. *Inf. Process. Manage.*, 41(5):1193–1205, September 2005.

[15] Abbe Mowshowitz and Akira Kawaguchi. Bias on the web. *Commun. ACM*, 45(9):56–60, September 2002.

[16] Leif Azzopardi and Ciaran Owens. Search engine predilection towards news media providers. In *Proceedings of the 32nd Int'l ACM SIGIR Conf. on R&D in IR*, SIGIR '09, pages 774–775, New York, NY, USA, 2009. ACM.

[17] Francis Galton. Vox populi (the wisdom of crowds). *Nature*, 75(7):450–451, 1907.

[18] H.C. Knight. *A Comparison of the Reliability of Group and Individual Judgments*. 1921.

[19] K. Gordon. Group judgments in the field of lifted weights. *Journal of Experimental Psychology*, 7(5):398, 1924.

[20] Richard Bruce. Group judgments in the fields of lifted weights and visual discrimination. *The Journal of Psychology*, 1:117–121, 07 2010.

[21] Samuel F. Klugman. Group judgments for familiar and unfamiliar materials. *The Journal of General Psychology*, 32(1):103–110, 1945.

[22] Jack L Treynor. Market efficiency and the bean jar experiment. *Financial Analysts Journal*, 43(3):50–53, 1987.

[23] James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[24] Oren Bracha and Frank Pasquale. Federal search commission? access, fairness, and accountability on the law of search. *Cornell L. Rev.*, 93:1149, 2008.

[25] In Amanda Spink and Michael Zimmer, editors, *Web Search*, volume 14 of *Info. Science and Knowledge Management*. 2008.

[26] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. Search bias quantification: investigating political bias in social media and web search. *Information Retrieval Journal*, 22:188–227, 2018.

[27] Robert Epstein and Ronald E Robertson. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521, 2015.

[28] Lucas D. Introna and Helen Nissenbaum. Shaping the web: Why the politics of search engines matters. *Inf. Soc.*, 16:169–185, 2000.

[29] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International Conference on World Wide Web 7*, WWW7, page 107–117, NLD, 1998. Elsevier Science Publishers B. V.

[30] B Edelman and B Lockwood. Measuring bias in “organic” web search. *Unpublished manuscript*, 2011.

[31] Irving D. Lorge, D. Fox, J R Davitz, and Malcolm Brenner. A survey of studies contrasting the quality of group performance and individual performance, 1920-1957. *Psychological bulletin*, 55 6:337–72, 1958.

[32] Joseph P Simmons, Leif D Nelson, Jeff Galak, and Shane Frederick. Intuitive biases in choice versus estimation: Implications for the wisdom of crowds. *Journal of Consumer Research*, 38(1):1–15, 2011.

[33] Lev Muchnik, Sinan Aral, and Sean Taylor. Social influence bias: A randomized experiment. *Science (New York, N.Y.)*, 341:647–51, 08 2013.

[34] Stefan M. Herzog and Ralph Hertwig. The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, 20(2):231–237, 2009.

[35] Albert Mannes, Jack Soll, and Richard Larrick. The wisdom of small crowds. 10 2013.

[36] Albert E Mannes, Jack B Soll, and Richard P Larrick. The wisdom of select crowds. *Journal of personality and social psychology*, 107(2):276, 2014.

[37] Albert Kao and Iain Couzin. Decision accuracy in complex environments is often maximized by small group sizes. *Proceedings. Biological sciences / The Royal Society*, 281:20133305, 04 2014.

[38] Daniel G. Goldstein, Randolph Preston McAfee, and Siddharth Suri. The wisdom of smaller, smarter crowds. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, page 471–488, New York, NY, USA, 2014. Association for Computing Machinery.

[39] Stylianos Kampakis and Andreas Adamides. Using twitter to predict football outcomes. *ArXiv*, abs/1411.1243, 2014.

[40] Stefan M. Herzog and Ralph Hertwig. The wisdom of ignorant crowds: Predicting sport outcomes by mere recognition. *Judgment and Decision Making*, 6:58–72, 2011.

[41] Stylianos Kampakis and William Thomas. Using machine learning to predict the outcome of english county twenty over cricket matches. *arXiv: Machine Learning*, 2015.

[42] Shiladitya Sinha, Chris Dyer, Kevin Gimpel, and Noah A. Smith. Predicting the nfl using twitter. *ArXiv*, abs/1310.6998, 2013.

[43] Joseph P. Simmons, Leif D. Nelson, Jeff Galak, and Shane Frederick. Are crowds wise when predicting against point spreads? it depends on how you ask. *ACR North American Advances*, 2010.

[44] Robert P. Schumaker, A. Tomasz Jarmoszko, and Chester S. Labedz. Predicting wins and spread in the premier league using a sentiment analysis of twitter. *Decis. Support Syst.*, 88:76–84, 2016.

[45] Weihong Cai, D. S. Yu, Ziyu Wu, Xin Du, and Teng Zhou. A hybrid ensemble learning framework for basketball outcomes prediction. *Physica A-statistical Mechanics and Its Applications*, 528:121461, 2019.

[46] Raza Ul Mustafa, M. Saqib Nawaz, M. Ikram Ullah Lali, Tehseen Zia, and Waqar Mehmood. Predicting the cricket match outcome using crowd opinions on social networks: A comparative study of machine learning methods. *Malaysian Journal of Computer Science*, 30:63–76, 2017.

[47] T. Falas, A. Charitou, and C. Charalambous. The application of artificial neural networks in the prediction of earnings. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 6, pages 3629–3633 vol.6, 1994.

[48] Z. H. Khan, Tasnim Sharmin Alin, and Akter Hussain. Price prediction of share market using artificial neural network (ann). *International Journal of Computer Applications*, 22:42–47, 2011.

[49] Shawndra Hill and Noah Ready-Campbell. Expert stock picker: The wisdom of (experts in) crowds. *International Journal of Electronic Commerce*, 15:102 – 73, 2011.

[50] Qili Wang, Wei Xu, and Han Zheng. Combining the wisdom of crowds and technical analysis for financial market prediction using deep random subspace ensembles. *Neurocomputing*, 299:51–61, 2018.

[51] Russ Ray. Prediction markets and the financial "wisdom of crowds". *Journal of Behavioral Finance*, 7:2 – 4, 2006.

[52] Anasse Bari, Pantea Peidaee, Aniruddh Khera, Jianghao Zhu, and Hongting Chen. Predicting financial markets using the wisdom of crowds. *2019 IEEE 4th International Conference on Big Data Analytics (ICBDA)*, pages 334–340, 2019.

[53] Zhi Da and Xing Huang. Harnessing the wisdom of crowds. *Behavioral & Experimental Finance eJournal*, 2018.

[54] Lawrence Brown and Joshua Khavis. The reliability of crowdsourced earnings forecasts. *Social Science Research Network*, 2018.

[55] Leigh Adam Drogen and Vinesh Jha. Generating abnormal returns using crowdsourced earnings forecasts from estimize. *Econometric Modeling: Capital Markets - Asset Pricing eJournal*, 2013.

[56] Russell Jame, Rick Johnston, Stanimir Markov, and Michael C. Wolfe. The value of crowdsourced earnings forecasts. *Econometric Modeling: Capital Markets - Forecasting eJournal*, 2016.

[57] Kailong Chen, Tianqi Chen, Guoqing Zheng, Ou Jin, Enpeng Yao, and Yong Yu. Collaborative personalized tweet recommendation. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 661–670. ACM, 2012.

[58] Nazpar Yazdanfar and Alex Thomo. Link recommender: Collaborative-filtering for recommending urls to twitter users. *Procedia Computer Science*, 19:412–419, 2013.

[59] Tianxi Li, Yu Wu, and Yu Zhang. Twitter hash tag prediction algorithm. In *ICOMP?11-The 2011 International Conference on Internet Computing*, 2011.

[60] John Hannon, Mike Bennett, and Barry Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 199–206. ACM, 2010.

[61] Shawn O'Banion and Larry Birnbaum. Using explicit linguistic expressions of preference in social media to predict voting behavior. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 207–214. IEEE, 2013.

[62] Tauhid R Zaman, Ralf Herbrich, Jurgen Van Gael, and David Stern. Predicting information spreading in twitter. In *Workshop on computational social science and the wisdom of crowds, nips*, volume 104, pages 17599–601. Citeseer, 2010.

[63] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Snehal Patil, Alessandro Flammini, and Filippo Menczer. Truthy: mapping the spread of astroturf in microblog streams. In *Proceedings of the 20th international conference companion on World wide web*, pages 249–252. ACM, 2011.

[64] Aditi Gupta and Ponnurangam Kumaraguru. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, page 2. ACM, 2012.

[65] Xin Xia, Xiaohu Yang, Chao Wu, Shanping Li, and Linfeng Bao. Information credibility on twitter in emergency situation. In *Pacific-Asia Workshop on Intelligence and Security Informatics*, pages 45–59. Springer, 2012.

[66] Saptarshi Ghosh, Naveen Sharma, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 575–590. ACM, 2012.

[67] Eugene Agichtein, Eric Brill, Susan Dumais, and Robert Ragno. Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th Annual Int'l ACM SIGIR Conf. on R&D in IR*, SIGIR '06, pages 3–10, New York, NY, USA, 2006. ACM.

[68] Fabian Wauthier, Michael Jordan, and Nebojsa Jojic. Efficient ranking from pairwise comparisons. In *Proceedings of the 30th Int'l Conf. on ML*, pages 109–117, 2013.

[69] Louis Leon Thurstone. A law of comparative judgment. *Psychological Review*, 34:273–286, 1994.

[70] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. 1952.