ABSTRACT

THE EMOTIONS OF SCIENCE: USING SOCIAL MEDIA TO GAUGE PUBLIC EMOTIONS TOWARD RESEARCH TOPICS

Cole Freeman, M.S.
Department of Computer Science
Northern Illinois University, 2020
Hamed Alhoori, Director

Online and in the real world, communities are bonded together by emotional consensus around core issues. Emotional responses to scientific findings often play a pivotal role in these core issues. When there is too much diversity of opinion on topics of science, emotions flare up and give rise to conflict. This conflict threatens positive outcomes for research. Emotions have the power to shape how people process new information. They can color the public's understanding of science, motivate policy positions, even change lives. And yet little work has been done to evaluate the public's emotional response to science using quantitative methods. In this thesis, we use a dataset of responses to scholarly articles on Facebook to model and analyze emotions toward topics in science.

NORTHERN ILLINOIS UNIVERSITY
DE KALB, ILLINOIS

MAY 2020

THE EMOTIONS OF SCIENCE:
USING SOCIAL MEDIA TO GAUGE PUBLIC EMOTIONS TOWARD RESEARCH TOPICS

BY

COLE FREEMAN

A THESIS SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE
MASTER OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

Thesis Director:
Hamed Alhoori

ACKNOWLEDGEMENTS

I cannot possibly give enough thanks to my advisor, Dr. Hamed Alhoori who has given so much of his time, energy, and attention to me during my time at NIU and continues to provide an example of what a hardworking, conscientious researcher is who is always growing and striving to do better. I would also like to thank my committee members, Drs. Michael Papka and Shupei Yuan, whose advice and support in the process of planning and undertaking this thesis has been invaluable.

I would also like to thank my fellow students, whose conversations and help during the course of this project were invaluable. Mrinal Kanti Roy and Michele Fattoruso, who were involved early on in this project—your conversations and hard work set this project on firm foundations early on. Suvo Chatterjee, who listened to me stumble through the ideas presented here often late into the night, and contributed many valuable insights that moved this project forward. Murtuza Shahzad and Akhil Pandey, along with my other colleagues in the DATA lab and ddiLab—thank you for listening to many of the problems I encountered in my work this past year, and for giving me detailed and helpful suggestions, many of which are represented in these pages.

DEDICATION

For Debra, Curtis, and Raychel.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

LIST OF APPENDICES

Appendix Page

CHAPTER 1
INTRODUCTION

1.1 Background

The rate at which information is being shared and reacted to on social-media platforms is increasing year by year. In June 2019, 293,000 new posts and 510,000 new comments were made on Facebook every minute [1], which makes for nearly 422 million posts generated each day. These platforms have further accelerated the speed of content evaluation and feedback with features such as click-based reactions. These relatively new modes of interaction provide users with a quick and easy way to respond to content in a manner that is still personal and emotionally expressive. They have been made available on platforms such as Facebook, which expanded its reaction palette in February 2016, and LinkedIn, which increased the number of click-based reactions available to users in April 2019. Until now, however, very little research has been undertaken to advance our understanding of these novel features.

Click-based reactions offer obvious benefits to researchers in the fields of bibliometrics and alternative metrics (or altmetrics)—a growing area of interest that takes into consideration the dissemination of a research outcome via multiple social-media platforms [2, 3, 4, 5, 6]. Previous studies have used citations as the gold standard for understanding and predicting the influence that scholarly research has on the scientific community itself [7], but evaluating the emotional impact that work may have on society has remained largely untouched. Click-based reactions may provide a way to approach this question of how research is affecting society.

1.1.1 Understanding the Lack of Attention on Facebook

First it will be useful to ask the question: why are researchers focusing on social-media platforms such as Twitter or Mendeley instead of Facebook? The answer, as it turns out, comes in several parts, and has been the subject of some recent scholarship in the field of altmetrics [8]. First, a number of studies have shown that academic articles appear more frequently and are shared more widely on Twitter and Mendeley than on Facebook. In 2016, Erdt et al. [9] found that only 7.7% of research articles are shared on Facebook, a fraction of the 59.2% that are shared to Mendeley and 24.3% that make it onto Twitter. Other studies have confirmed this trend, such as Zahedi and Costas [10] where study of a new dataset suggested that somewhere between 7.9% and 16.3% of research articles make it onto Facebook.

A second reason less attention has been paid to Facebook is that there has been a consensus in the scholarly community over the past few years that specialists in scientific and academic fields tend to prefer Twitter over Facebook for micro-blogging, sharing content, or interacting with others. In a study published in 2016, Collins et al. [11] found that 93.7% of academics who participated in a survey said that they had a Twitter account, with a margin of error of ±2.03% at 95% confidence. They also found that 87.7% of respondents used Facebook, with a margin of error of ±2.75% at 95% confidence. Overall, the study shows a slight but significant preference of Twitter by experts. These numbers are displayed in the upper two bars of Figure 1.1, in which we see a slight preference of Twitter over Facebook in a sample of academic researchers.

A third reason is that Facebook is more protective of their data than other platforms such as Twitter [8]. In the wake of controversies surrounding the 2016 U.S. elections, Facebook has enacted a series of more and more stringent policies to keep outside parties from collecting

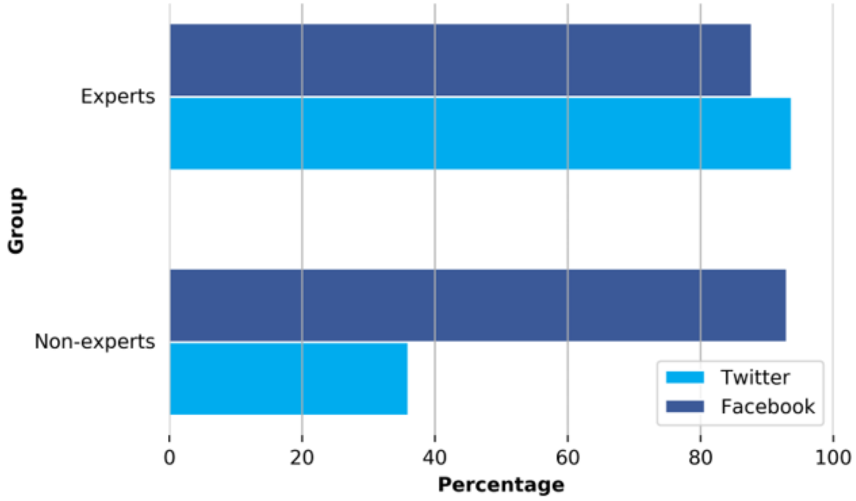


Figure 1.1: The disparity between expert and non-expert usage of Twitter and Facebook.

information about users or content on the website. After data which was ostensibly gathered for research was used for political purposes, the company became more selective about who they gave API access to. Applying for access to research Facebook content involves filling out copious amounts of forms, long periods of waiting, and sometimes financial resources that are in all quite restrictive and uninviting—all this while a relatively simple web-scraper can access Twitter data with relative ease. In short, Facebook data is hard to get at. We experienced this first hand in our study. Even with proper access, we had to be creative with our query methods so as not to be blocked by their Graph API for collecting too much data, or for collecting data too quickly, or for trying to automate parts of our data collection.

1.1.2 Why Focus on Facebook?

These three factors have worked together and resulted in the platform being relatively understudied in the past few years. So why did we want to see what was happening with shares of research on Facebook? Again there are several answers to this question.

Though research is shared more consistently on Twitter than on Facebook, patterns of sharing differ between the platforms. A 2015 study by Costas, Zahedi, and Wouters [2] found that health sciences and biomedical research were the most popular areas of research to make its way across Facebook, a pattern of sharing that was not observed on Twitter. Social-media coverage of research also differed from country to country. A 2015 paper by Alperin [12] found that Brazilian researchers were more active on Facebook than researchers from other countries, and that Latin American journals saw more coverage on that platform than did journals from other countries. So we conclude that studying Facebook might be beneficial depending on what groups of people or fields of study a researcher is interested in.

Another reason is that though Twitter may be the platform of choice for experts, this is clearly not the case among the general population. The same study by Collins et al. [11] that found a disparity between academics' usage of various online platforms found that 93% of non-expert social-media users are on Facebook, while only 36% are active on Twitter. This disparity of social-media usage between groups is displayed in Figure 1.1. This wide gap between expert and non-expert use of platforms suggests that social-media research up to this point has largely followed expert use. This is a fine way of using social-media data, but it is important to note that it can only answer certain kinds of research questions we might want to ask. To answer other questions—for example, what are non-experts saying about science online—we would have to change the types of data we are looking at.

Another reason to look into Facebook is the sheer number of users it has, as well as the rate at which content is being proliferated there. In March of 2019, Facebook reported that it had 2.38 billion active monthly users—–in other words about 31% of the world's population is on Facebook. This dwarfs other platforms, such as Twitter, which reported 126 million users in February of 2019, or Instagram, which had roughly a billion active users, and LinkedIn, which had about 500 million active users. Figure 1.2 compares the number

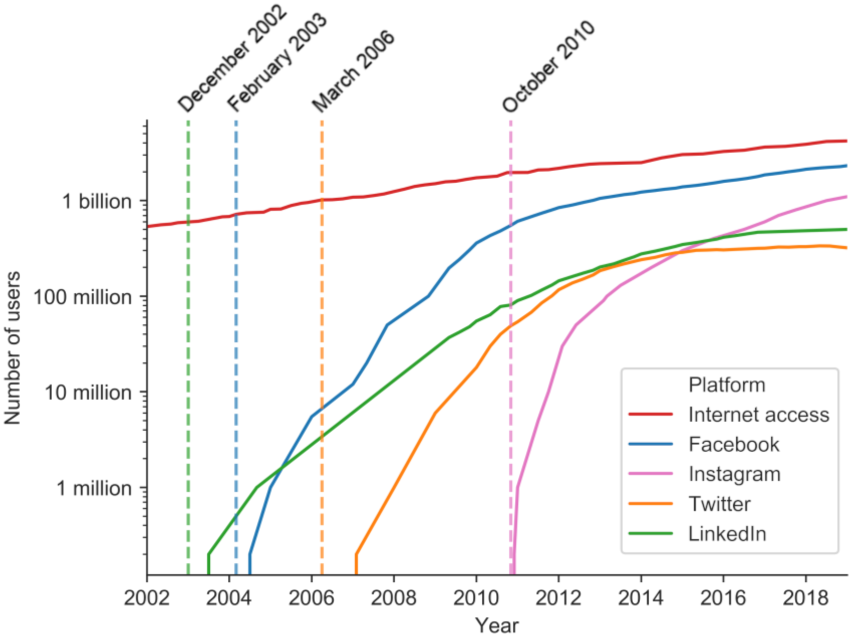


Figure 1.2: Comparison of the number of users on four of the most prominent social-media platforms, with the number of internet users world-wide displayed as well. The y-axis is scaled logarithmically.

of active users reported on four of the most prominent social-media platforms from their inception to 2019. It also shows the number of people world-wide with internet access.

Part of what motivated our project was these questions. We were driven by a curiosity to discover what's happening on Facebook that we might otherwise miss. What can we do with Facebook data that we cannot do with data from other platforms? Just because it may be difficult to access data, that does not mean we should be deterred. We were also motivated by broader research questions about public understanding and engagement with science. Learning how the public is interacting with and emotionally responding to scientific findings could provide important insights for science communicators in a broad range of disciplines. It could help scientists and research institutions fine-tune their public messaging strategies for maximum effect; it could help researchers better appreciate when their findings are not fully understood by the public, or when they are generating unexpected emotional reactions; it could also help researchers identify when they are working in directions that are considered controversial by lay-persons, aiding them in preemptively addressing issues that may have negative outcomes when their findings are published or communicated to the public.

It is often the case that advancements in scientific understanding are preceded by advancements in measurement. Though there are well-established methods and tools for analyzing the sentiment of texts, the novelty of click-based reactions has meant that there are not the same resources available for analyzing emotions expressed through clicks. In this thesis we present several studies that turn to public pages of science on Facebook for insights about the public's understanding and response to research. We survey the pages themselves, learning what we can about the communities that develop around scientific findings on that platform and their motivations for engaging with research outcomes. We develop a heuristic for analyzing click-based responses to social-media content. We borrow from psychological research the concepts of emotional valence, intensity, and diversity, using them to gain an improved understanding of how Facebook users are responding to scientific research through their reactions. We develop metrics for these emotional concepts after careful analysis of a dataset of Facebook reactions that we collected for this study. We apply our newly developed methods and metrics to our data and perform statistical tests to identify significant trends. We also use statistical modeling to learn about the public's emotional engagement with topics of science. We present and analyze the results of these models in several ways, trying to gain a better understanding of the patterns of emotional response that are present in our data. We are most interested in learning about aggregate behavior and sentiment toward topics of science, and the models and methods we use find different ways of approaching this problem using our large and rich dataset of click-based reactions.

1.2 Overview

This section presents the format that the subsequent chapters will take. We give salient details for each chapter in our thesis, and then present some of the primary contributions our thesis makes toward computational social science research and altmetrics.

1.2.1 Outline

Chapter 2 gives a background of the scholarly work that has used Facebook reactions and other social-media data to conduct research tasks related to sentiment analysis, for health-related studies, and other related research areas. It gives an overview of the data we work with in our studies, where it came from, and some of the techniques we used to collect it.

In Chapter 3, we explore the data we collected. We focus on the public Facebook pages where scientific research is being shared, analyzing common themes between pages, and identifying some interesting practices and uses of science on the social-media platform. We also identify some outliers in our collected data that reveal unusual, pattern-breaking uses of click-based reactions.

Chapter 4 uses our collected data to develop new metrics for measuring emotional responses to content based on click-based reactions. We present these metrics, which are derived from principles in social psychology, and apply them to our data to show interesting patterns of group behavior. We use our metrics to perform some hypothesis testing. The results are analyzed and interpreted to reveal more about our data and the tools we use to explore it.

The final study is presented in Chapter 5. We use a Latent Dirichlet Allocation (LDA) topic model to represent our documents as a feature vector. We use our click-based reactions as categorical responses to the documents as topic vectors to train a multinomial logistic regression model. We analyze the coefficients of our model to learn on average how increases in the presence of a scientific topic motivate specific emotional responses from users.

Figure 1.3 shows the overall structure of our study, from data to experimentation. Each study builds of the knowledge gained in the previous studies, using the results to motivate further questions in the next. The studies are also meant to develop in complexity. The first uses a primarily qualitative approach with some statistical analysis. The second study develops metrics based on common practices in information retrieval such as term frequency-inverse document frequency, as well as methods from information theory, such as the Jensen-Shannon distance metric and entropy. The final study takes a sophisticated approach to LDA topic modeling, and utilizes relatively novel techniques from linear algebra and statistical analysis to build logistic regression models with compositional data.

1.2.2 Contributions

This thesis offers several contributions. First and foremost, it offers an extensive dataset of shares of scientific articles on Facebook, along with the click-based reactions that these posts received. This dataset has already been utilized for several studies by different researchers, and it is our hope that it can be useful for researchers in a variety of disciplines for years to come. We also offer several metrics, tools, and methods that might aid researchers who are interested in performing similar studies. In Chapter 4, we develop metrics for emotional diversity, intensity, and valence using click-based reactions, as well as a novel method for transforming these reactions using the frequency of their overall usage.

In Chapter 5 we present a method of using compositional data (data that sums to a constant in every row, such as proportional data or percentages) as the input for regression

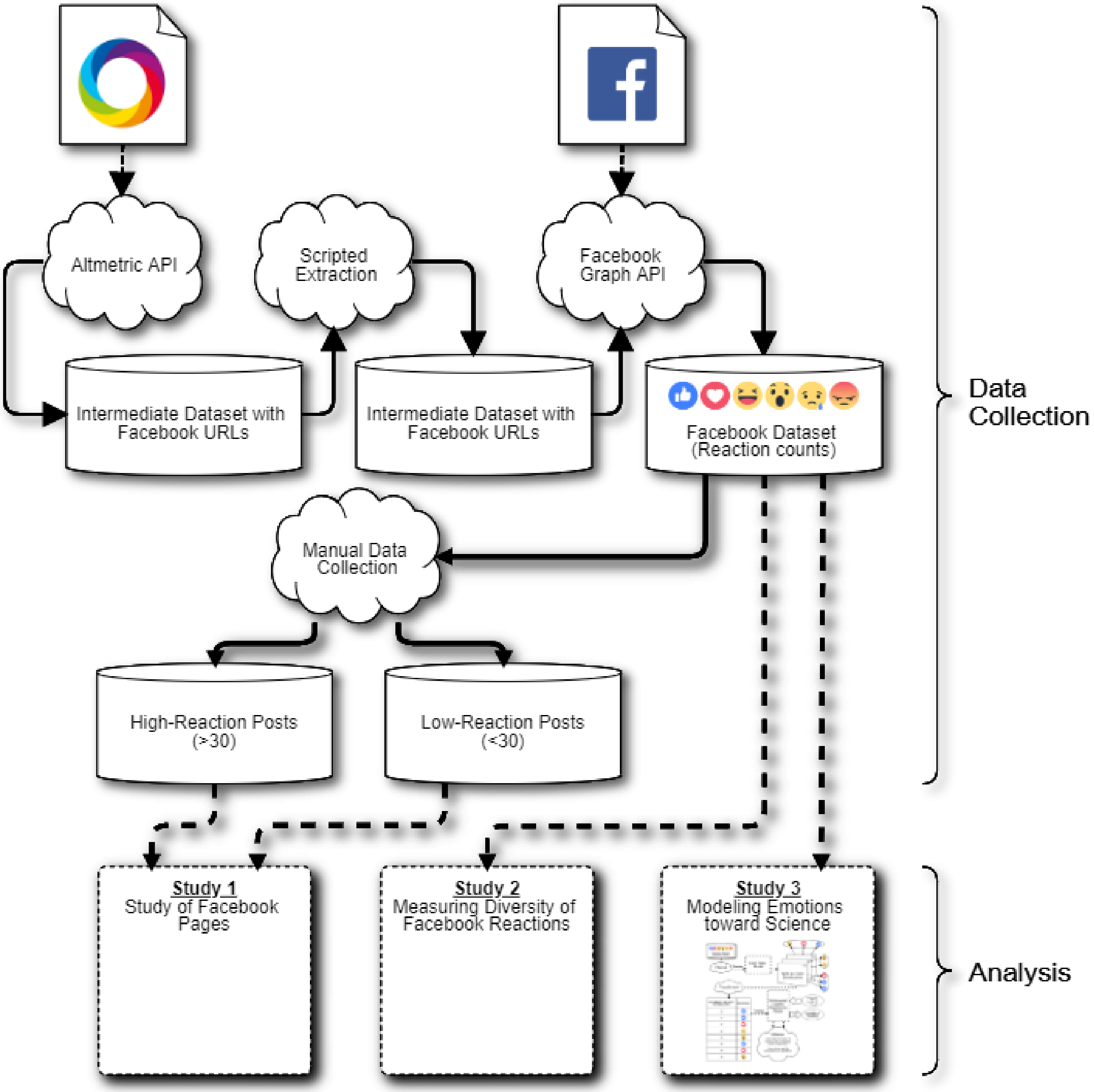

Altmetric API
Scripted Extraction
Facebook Graph API
Intermediate Dataset with Facebook URLs
Intermediate Dataset with Facebook URLs
Facebook Dataset (Reaction counts)
Data Collection
Manual Data Collection
High-Reaction Posts (>30)
Low-Reaction Posts (<30)
Study 1
Study of Facebook Pages
Study 2
Measuring Diversity of Facebook Reactions
Study 3
Modeling Emotions toward Science
Analysis

Figure 1.3: Outline of this project, from data collection methods to analysis.

models that leans on statistical methods that are not common practice in machine learning. We develop several Python classes that handle compositional data. We hope that the work performed in this study will inspire other researchers in machine learning and statistical analysis to use our tools. We also show novel and convincing evidence of patterns of emotional reaction to topics in science on social-media platforms. Social-media data is known to be noisy and difficult to model with. By properly cleaning and transforming our large dataset, we are able to show patterns of emotional response that are meaningful and can help scientists and science journalists target their studies and communications with the public for maximum effectiveness.

CHAPTER 2

BACKGROUND

2.1 Term Definitions

We begin by defining a few important terms:

- **Click-based reactions**: features on social-media platforms that allow users to leave a quick and easy response to content; click-based reactions are non-textual and are related to textual emojis, which convey emotional responses through pictograms; on Facebook, click-based reactions include the six buttons “Like,” “Love,” “Wow,” “Laughter,” “Sad,” and “Anger”; Figure 2.1 shows the six click-based reactions from Facebook.

- **Five special reactions**: the five click-based reactions: “Love,” “Wow,” “Laughter,” “Sad,” and “Anger.”

- **Page visibility**: the number of followers a Facebook page has; Facebook allows users to like or follow pages; after following a page, a user will begin to see content shared by or on those pages on their own timeline.

- **Shares**: the number of public Facebook pages an article has been posted onto.

- **Reshares**: the number of times users have re-shared a public post of an article onto another private or public page. Our dataset contains reactions to the initial share and the number of “Reshares,” but we do not have reaction data from article “Reshares.”

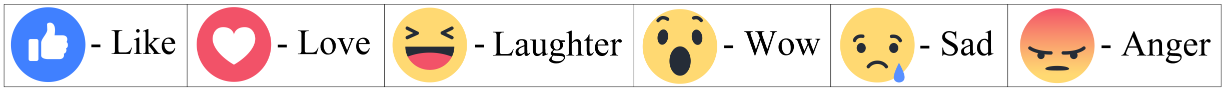


Figure 2.1: The six click-based reactions available to users on Facebook.

- **Articles**: our dataset consists of Facebook reactions to research related articles; for convenience, these are sometimes referred to as "documents," especially in the description of feature transformation.

2.2 Measuring Emotion

When evaluating the emotional responses that a group of users have to social-media content, three factors are particularly relevant: the *valence*, *intensity*, and *diversity* of the response. In psychological literature, these factors usually refer to the emotional response of an individual [13, 14, 15, 16, 17]. Since we are interested in measuring aggregate responses here, we are instead treating these three variables as indications of aggregate response to content. Taking measurements of individual users would not be possible with the data we collected for a number of reasons. First and foremost, we took precautions to avoid collecting information that could be used to identify specific users; measuring more granular emotional responses would necessarily require more information about the individuals. Secondly, each user can only click one reaction on a given post (they can, however, share and provide a reaction to a post). Although this information is useful in finding group responses, it is not at all helpful for identifying when an individual has a diverse or conflicted response to a given paper.

The valence of response denotes the "direction" of the emotion (i.e., positive or negative). Positive emotions promote social connections and playfulness [14], whereas negative emotions promote wariness and avoidance, though they sometimes capture more attention than positive emotions [13, 18]. Valence is a binary measurement, and it is important to note that negative responses can have beneficial effects for both individuals and communities [18]. The intensity of an emotional response reflects the importance of information to respondents [15, 19, 20]. As the intensity of emotion increases, the object of the emotion captures more of individuals' attention [16, 13, 21, 22]. Intensity can also be used as an indicator of how long information will be retained in a person's memory [18]. More intense emotions are also more likely to influence decision making [23].

The final factor we use here is diversity of response. It indicates the degree to which the user response to a post simultaneously point in different emotional directions. It can indicate that content is controversial or that there is not consensus about how a subject should be received or interpreted by a group of responders. A higher diversity of emotional response about a post or article indicates that there is much variation in individuals' responses and that there is not consensus about a particular subject or issue [17].

2.3 Topic Models

A topic model is a category of statistical model used in Natural Language Processing (NLP) to discover latent semantic structures or significant word groupings that occur in a corpus of texts. They are often built with a large body of documents as a way to preserve the distinct properties of each document while also giving a brief description of each as a unique mixture of topics. Topic models facilitate useful basic tasks in machine learning, such as summarization of texts, determination of document relevance, detection of novelty, as well as classification.

Latent Dirichlet allocation (LDA) topic models were first applied in a machine learning context by Blei et al. [24]. The underlying idea of LDA is that documents can be represented as “random mixtures over latent topics,” and that each topic can be represented as a probability distribution over words. It is this type of topic model that has been the most widely applied in text mining literature. An LDA model takes as input a corpus of documents and a given number of topics t that the user would like to discover, and defines t groups of words or n-grams that frequently co-occur. Each document can then be represented as a distribution over the t topics.

2.4 Literature Review

Studies in social-media analytics tend to focus on text, using approaches such as NLP, sentiment analysis, opinion mining, or graph mining to arrive at and support research conclusions, or on the proliferation of content through online communities [25, 26, 27, 28]. These approaches have proved effective for understanding and predicting many aspects of human behavior, but they leave a number of other expressive signals unexamined. Click-based reactions, on the other hand, are a relatively underutilized resource in social-media research. Examples of quick-draw, ready-made expressive features are becoming increasingly prevalent across many platforms, and as such have attracted some amount of attention from researchers in the past few years.

Several studies have been conducted to measure and understand emotions using social-media data [29, 30, 31, 32, 33, 34]. Tian et al. [35] studied the way Facebook users modify the sentiment of their comments with emojis. They targeted posts on public news pages, comparing three channels of expression: natural language, emojis, and reactions. They found that generally the emotional content of emojis and reactions correspond, but that in instances of sarcasm or politeness, the two channels could express different meanings. Krebs et al. [36] used customer satisfaction data gathered from Facebook to train a model using convolutional and recurrent neural networks (CNN and RNN) and predict the reaction distribution for a given post. Basile et al. [37] combine NLP and sentiment analysis of Facebook reactions to build a regression model that predicts news controversies in Italian media.

There have been several studies on community building, social interaction, and identity on Facebook and other social-media platforms. Rohde et al. [38] and Hewitt et al. [39] were early examples of how social-media data could be used to study inter-community interaction and identity formation online. Burke et al. [40] found that Facebook users are more likely to respond with greater emotional intensity in both positive and negative directions to community members' posts if their friend networks are smaller and more greatly connected. Thagard and Kroon [41] highlighted the role emotional consensus plays in group decision making community cohesion. Their conception of emotional consensus was built on the idea that groups reach accord when all party members communicate the valences they associate with each possibility. The authors argued that consensus is reached in part through rational discussion and argument, but that we overemphasize the importance of this because we tend to believe that humans behave as rational actors. They looked to psychological research to show that group behavior is driven much more by nonverbal forms of communication such as "facial expressions, voices, postures, [and] movements." In working toward consensus, emotionally stronger positions are "contagious," spreading through the group and finally having the greatest influence on outcomes.

Kumar et al. [42] look at conflict and confrontations between communities on Reddit. They define communities by the users who participate in distinct forums on the site (i.e., "subreddits"), each of which caters to specific interests of users and are curated by page moderators. This definition of community as a well-defined space where members frequently visit and interact with other members works well within the structure of Reddit. Reza et al. [43] label this type of community where users understand that they are a member of a community and interact with other members of their community more than nonmembers as *explicit*.

Participation on Facebook can be described as explicit, but the user experience there is not necessarily centered around community pages, but rather around a kind of "commons" area—the news feed—where users interact with family and friends and see content that algorithms predict will be of interest to them. Facebook members define their interests and then receive content catered to their wants. The result is that a single user can follow and interact with content on many different pages, often without even directly moving onto the pages from which the content originates. Reza et al. [43] refer to participation in an unacknowledged community as *implicit*.

Some studies have shown that a person's emotions, such as anger, sadness, happiness, and depression, differ from each other in terms of the extent of their influence on other people. Rosenquist et al. [44] used a longitudinal statistical model to analyze a social network of 12,067 people from the Framingham Heart Study [45], a long-term, ongoing cardiovascular cohort study of residents of the city of Framingham, MA, to determine whether symptoms of depression in a person are associated with their friends, co-workers, siblings, spouses, and neighbors. To assess depressive symptoms, the researchers used the Center for Epidemiological Scale. The results showed that an association can be found in people at up to three degrees of separation—i.e., from the depressive person's friends to their friends and then to their friends. Researchers have also studied variations of these emotions as shown by online users. Fan et al. [46] used a multi-emotions classification model pertaining to anger, joy, disgust, and sadness to determine how these emotions correlate on Weibo—a Chinese microblogging website. Using both Pearson correlation and Spearman correlation, they found that different sentiments have different correlations. Their study showed that correlation among users are high for anger and low for sad sentiments.

Other studies such as Vimala et al. [47] have used Facebook data to study health-related issues. In their work, they use Facebook communities centered around diabetes to detect and classify emotional responses, examining the relationship between textual expressions of emotion and click-based reactions. They build a predictive model from their findings, using it predict group emotional states based on the presence of certain reaction types with promising results.

Burnap et al. [48] not only included a sentiment analysis but also details pertaining to users' prior party support to predict the outcome of the 2015 UK general election. Having trained a model using nearly 14 million tweets and relying on a range from extremely negative to extremely positive to describe users' sentiments, the researchers predicted that the Labour Party would win the election. In a similar study, Vepsalainen et al. [49] examined how Facebook "Likes" can be used to predict election outcomes. They collected 2.7 million data-points using Facebook's Graph API and used Absolute Error to measure the accuracy of their predictions. In that study, the authors were surprised to find that "Likes" were not a strong indicator of the election outcome. They look into the reasons they may have achieved this result: skewed demographics, uneven activity by the candidates in social media, and noise in the sample.

Several studies have been conducted to understand altmetrics and the societal impact of research [50, 51, 52, 53, 54]. Recently, researchers have used online metrics to predict citations in news outlets [55], public policy documents [56, 57], and patents [58].

We propose to use this wealth of information about the social and media dissemination of research as an indicator of societal impact, as it reflects not only immediate interest in a research finding, but also the degree to which individuals find the work to be of sufficient interest to warrant sharing. Additionally, the value of understanding the public's interest and reactions help science communicators identify effective ways to engage with the public and build positive connections between scientific communities and the public.

These studies helped us to look at the challenges of understanding and measuring the public engagement with research from several perspectives, to better fathom this complex problem, and to provide a foundation for the proposed project. To the best of our knowledge, this is the first large-scale study to understand and measure the emotional impact of science.

2.5 The Datasets

2.5.1 Dataset A

Our dataset is made up of articles discovered through Altmetric's online database.[1] Their database holds information about millions of scholarly articles, research studies, and news about scientific findings published in a variety of languages and disciplines. The articles in our dataset were retrieved from a data dump collected in July 2018. We filtered the articles we were targeting to only those that had been shared on a public Facebook page one or more times. We further filtered the articles to only those published in 2017. Choosing this year accomplished three goals. (i) Reactions were released by Facebook in February 2016 [59], so any articles we looked at had to be published after that time to have meaningful data on this feature. (ii) Whenever a new feature is rolled out, it takes time for users to learn how to use it; Shah [60] finds that use of reactions increased from 2.4% of all interactions in April 2016 to 5.8% by June 2016, and up to 12.8% of all interactions by June 2018; by the time of our data collection in early 2019, a large enough subset of users were comfortable expressing themselves with the feature to warrant more scholarly attention. (iii) By the time we began our data collection, a sufficient interval of time had passed for articles to be widely shared and reacted to (between 15 and 30 months).

[1] https://www.altmetric.com/

Altmetric's database provides URLs for shares of its articles onto public Facebook pages. We used these links to query Facebook's Graph API for information about user reactions to the posts. This process is limited to 200 queries per hour, with each individual query retrieving (i) the click-based reactions, (ii) the number of "Reshares" that post received, and (ii) any text included with the share for one share of an article onto a public page. Some articles are shared many times to many different pages; for these, we collected the information for each post and summed all of the reactions into a total reaction score. We did not collect information on the comments added to the posts by users, nor on the article "Reshares." The range of article shares in our dataset was between 1 and 362. The median number of shares that articles in our dataset received was 1. The mean number of shares was 2.30 with a standard deviation of 4.57. Clearly, the distribution of article shares is skewed right—a minority of articles received many times the average number of shares.

We collected data on 356,664 shares of 149,747 scientific articles. Most of the collecting was undertaken between March and July 2019. For this study, however, we were interested in exploring how Facebook users are employing click-based reactions; we thus limit the articles we were looking at to only those that had received one or more of the five special reactions. Our final filtered dataset included 33,662 articles shared onto 178,403 public Facebook pages, all of which received a total of 6,418,053 click-based reactions and 2,051,299 "Reshares."

For each article, our dataset includes: article title, article abstract, article publication date, the number of public Facebook pages the article was shared onto, and the number of click-based reactions of each category. It also records the text, if any, that was included along with the post. There are also several features that record article topics: "Subjects," which includes the subject areas pertinent to each article that were selected by the authors; "Scopus subjects," which is the subject areas that are recorded in each article's entry in the Scopus database; and "Publisher subjects," which records the subject areas of the journals in which the articles are published. Each of these features can contain one, many, or no subject(s) per article.

In our data collection process, we took the utmost care to respect Altmetric's and Facebook's specifications for how and why their data can be accessed and used. We prioritized avoidance of collecting personally identifying information regarding specific social-media users. Our interests were only in the ways people interact as a whole with scholarly content on social-media platforms, not in how specific user's beliefs or opinions may influence their behavior. We recognize that identifying information could in some instances be inferred *a posteriori* from some of the data we collect; however, our method of data collection does not target anything that could be used to consistently identify individual users.

2.5.2 Dataset B1 and B2

There were several questions we had that our original dataset could not help us answer: What are the Facebook pages that these research articles are being shared onto? How can we contextualize the quantitative data we collected? Can a sample of the public pages where scientific articles are being shared help us better understand emotions toward science as displayed on Facebook? These and other questions arose as we studied Dataset A; we therefore decided to collect supplementary data that could help us address these problems.

As described above, we collected Dataset A through an automated process of queries to Altmetric's and Facebook's APIs. Throughout the process, our job as researchers was (1) to create a script that properly queried these APIs and then catalogued the results in a safe and orderly format that was stored on local devices, and (2) then to ensure that no problems came up during the collection process, and if any problems did occur to address them appropriately. These queries only gave us access to information about the posts themselves, but not about the pages onto which they were shared. To remedy this shortcoming, we selected a random sample of 200 highly-reacted-to posts—which we defined as posts that received more than thirty combined special reactions—from Dataset A and manually visited each post and the public page onto which it was shared. We recorded important details about both the post and the page in a new dataset.

The features we collected are listed and described in Table 2.1. Many of these features are analogous to the features in Dataset A, but some hold information that was not collected before. Whether or not a feature is unique to Datasets B1 and B2 or is shared between them and Dataset A is denoted by the third column in the table. One particularly interesting different feature is the “Other” category of reactions. These were special event reactions Facebook released for a limited amount of time. Our original data collection did not target these categories, and therefore they do not show up in Dataset A. The number of posts that had these special events reactions were relatively small, and there were not many applied to any given post.

We selected the target posts by filtering Dataset A to only those posts that received more than thirty total special reactions across all article shares. We then randomly selected 200 of these articles. For each of the posts of these 200 articles that individually received more than thirty reactions, we appended the post URL into a list. We followed these links to the pages, copying the target features from the page into Dataset B1. Each article can be shared onto multiple pages and sometimes even to the same page multiple times. We added any post that received more than the thirty-reaction threshold. The result was a set of 224 Facebook posts shared onto 122 unique public pages. We collected Dataset B1 between January 20-22, 2020. Since some of the posts were made onto the same public page, we maintained consistency of the data by copying and pasting the details about the public page from the first time we encountered it. Many of the features of the posts and pages are dynamic (e.g., the number of reactions a post has received, and the number of followers a page has), and so to avoid

Table 2.1: Features of Dataset B1 and B2.

Feature	***Description***	***New Feature***
Altmetric ID	Unique identifier for article in Altmetric database	N
Post URL	Hyperlink to Facebook post	N
Post Text	Text included on post (if any)	N
Comment Count	The number of comments added to post	N
Share Count	The number of times the post was reshared by users	N
Video	Binary feature (y/n) indicating whether post includes a video	Y
Video Views	If post is a video, the number of times the video was viewed	Y
Likes	The number of “Likes” the post received	N
Loves	The number of “Love” reactions the post received	N
Wow	The number of “Wow” reactions the post received	N
Laughter	The number of “Laughter” reactions the post received	N
Angry	The number of “Angry” reactions the post received	N
Sad	The number of “Sad” reactions the post received	N
Other	The number of other reactions the post received (special event reactions: rainbow or flower reactions)	Y
Page Name	The name of the public Facebook page	Y
About	Text included in the **About** section on the Facebook page	Y
Community Standards	Binary feature (y/n) indicating whether the page has rules that followers are expected to abide by	Y
Page Likes	The number of people who have “Liked” the public page	N
Page Follows	The number of followers the public page has	N
Check-ins	The number of in-person check ins a page has been (where a user has physically visited the organization managing the Facebook page)	Y
Page Type	Categories of the Facebook page; selected by the page owner and displayed in the **About** page	Y
Misc.	Miscellaneous information about the page (e.g., if there are notable features that distinguish it from other pages)	Y
Language	Language of the page	Y
Person	Binary feature (y/n) indicating whether the page represents an individual person or not	Y

Table 2.2: Description of the three datasets.

Dataset	**Selection**	**Dates Collected**	**No. of Posts**	**Collection Method**
A	(all)	Mar.–Jul. 2019	356,664	Automated
B1	reactions > 30	Jan. 20–22, 2020	224	Manual
B2	reactions ≤ 30	Jan. 27–29, 2020	171	Manual

inconsistencies such as a single page having different follower counts from one instance to the next, we chose to record the information encountered at the first visit to each page.

Though we were primarily interested in posts that garnered a significantly large amount of attention, which we defined as those that received more than 30 of the five special reactions, we also wanted to explore the pages with posts that did not reach this threshold. Dataset B2 collected the same features as those in B1, but we limited the posts used to access pages by those that received no more than 30 of the five special reactions. We then randomly selected a group of these posts and followed the same methods outlined above. Dataset B2 contains 171 posts on 141 unique public Facebook pages. It was collected by manually accessing the pages between January 27 – 29, 2020. Table 2.2 shows significant information that distinguishes our three datasets from each other.

CHAPTER 3

PUBLIC PAGES OF SCIENCE ON FACEBOOK

> *What brings men together is not a community of views but a consanguinity of minds.*
>
> Marcel Proust

In this first study of our data, we are interested in performing a qualitative exploration of the features of some of the public pages of science where research is being shared on Facebook. We want to discover patterns information sharing, as well as to see what motivates people to share research conclusions on social-media platforms such as Facebook. We want to answer questions such as: What are the goals that content sharers have? and how do content sharers present themselves on these public pages? These patterns will help us interpret the results of the studies presented in Chapters 4 and 5. They will provide a valuable context for some for the results we see with later analyses, and will give us greater power to interpret some of the behavior we identify in our data. This chapter will also serve as an introduction to a subset of our data, giving the reader a better idea of what kind of posts we collected in our dataset and what features we are interested in studying.

Public pages of science on Facebook are varied in structure, in the kind of content they produce and share, and in their approach toward fostering a communal space for interactions between users; and no single set of pages could completely capture the diversity that exists in this space. We did, however, find several features that serve as a basis for comparison between pages, and we were able to use this basis to identify trends of information sharing across pages. We begin by surveying some of the main types of pages we found.

3.1 “We’re here to talk about ideas”

Overall, we find that the creators and moderators of the public Facebook pages of science in our datasets want to foster an environment where good information is shared, and people can come to have their ideas challenged but not attacked. There are of course pages that do not fit this description, and they will be discussed as well. What brings users to a given page and keeps them interacting with content appears to be what Marcel Proust described in a quote from *Remembrance of Things Past* [61] included as the epigram to this chapter: common points of view are not sufficient on their own to produce a sustained community—it takes a kind of similar mindset or a shared feeling about the importance of the underlying subject matter to allow a group of individuals to cohere. One of the pages we studied contained the following mantra, which seems to encapsulate much of the information we saw across pages: “We’re here to talk about ideas” [62]. The largest pages that appeared to receive the most attention and interaction seemed to be those that promoted a sense of the importance of good information about a range of topics, or a sense that the specific topic dealt with by a page is of global importance.

3.1.1 Types of Pages

Figure 3.1 shows several different page types we discovered in exploring pages of science on Facebook. We group the pages into three broad categories: pages representing scientific publications, general scientific information pages, and politically oriented pages. Each of these are broken into several subcategories of pages. These categories obviously have some overlap. For example, some general information pages of health science will share political posts from time to time.

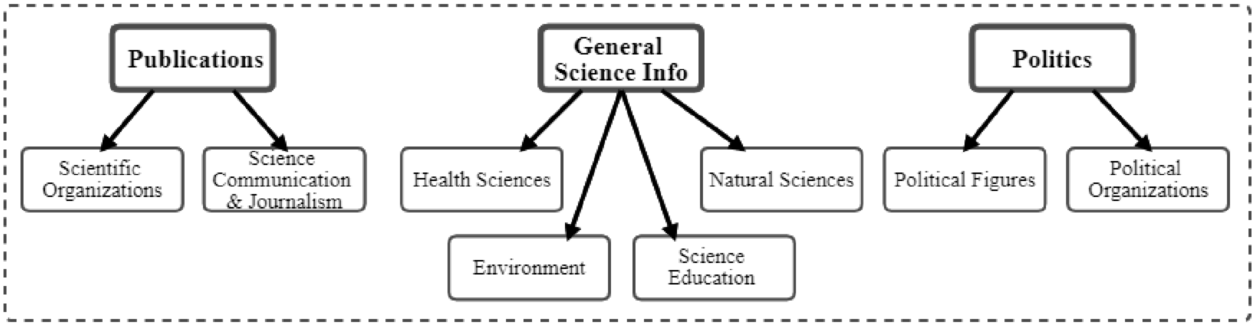


Figure 3.1: Several Facebook page types that shared scientific articles

Table 3.1: Ten Facebook pages with the most shares in Datasets B1 and B2, with the number of followers on each page.

Facebook Page Name	**Num. of Posts**	**Num. Followers**
The Conversation	17	273,771
Neuroscience News and Research	15	1,718,739
Nature	10	984,612
New England Journal of Medicine	8	1,764,493
Scientific American	8	3,188,329
News from Science	7	2,429,124
Chemistry World	6	968,038
The Science Explorer	6	1,678,881
Science	5	4,072,094
arXiv-Social	5	1,111

3.1.1.1 Large Publication Pages

The largest share of posts in Dataset B1 and B2 were made by pages that represented popular publications and science content providers. Table 3.1 shows the ten pages with the most posts, as well as the number of followers on each of those pages. Three of the ten pages represent specific disciplines (“Neuroscience News and Research” covers topics from biology, neurology, and sometimes even psychology, “New England Journal of Medicine” covers topics from health sciences, and “Chemistry World” covers topics in chemistry and biology) while the other seven share articles across a broad range of disciplines.

“The Conversation” [63] appeared the most frequently in Datasets B1 and B2, representing 17 posts, even though it had a significantly smaller number of followers than some of the other pages posting about research. Its “About” section gives a clear and concise motivation for the page: “The Conversation is an independent channel of information, analysis and opinion – sourced from the rich expertise of the university and research sector.” According to their website, The Conversation “arose out of deep-seated concerns for the fading quality of our public discourse—and recognition of the vital role that academic experts can play in the public arena” [64].

“Food Science and Nutrition” [65] has 459 thousand followers, and represents one facet of the global publishing company Elsevier. Posts on that page are mostly related to material published in their journals to nutritional, agricultural, and biological science, as well as to remind followers of deadlines for article submissions and to share general information about their publications.

3.1.1.2 Scientific Organizations

Some of the pages are devoted to specific disciplines of science and academia such as biology, physics, or environmental and health sciences. “Chemistry World” [66] offers to keep its nearly 1 million followers up to date with the “stories from across the chemical sciences.” It also offers “science news, research, reviews, features and opinions” about findings across a variety of scientific fields. It posts about biology, academia, even the arts—an example of the last category is a post made on February 26, 2020 about an annual award that “challenges scientists to communicate their doctoral research using interpretive dance.”

Other pages represent organizations or professional societies for scientists and researchers. The “American Society for Microbiology (ASM)” [67] has about 375 thousand followers, and posts about recent discoveries in the field of microbiology, as well as about conferences and publications organized by the ASM. Their "About" section outlines the history of the organization, and its importance in its field: "The ASM is the oldest and largest single life science membership organization in the world. Membership has grown from 59 scientists in 1899 to more than 39,000 members today, with more than one third located outside the United States. The members represent 26 disciplines of microbiological specialization plus a division for microbiology educators." Some of these organizations are more general, such as "The Nobel Prize," [68] which is the official page for that organization. It has nearly 4.5 million followers and shares information about a broad range of scientific domains. Most of their posts relate to the work of Nobel laureates, or are about the social or scientific contributions of laureates.

The page for "New England Journal of Medicine" [69] states that they are "dedicated to bringing physicians the best research and key information at the intersection of biomedical science and clinical practice, and to presenting the information in an understandable and clinically useful format." Their page shares content focused narrowly on health sciences.

3.1.1.3 Political Pages

Political activism was a primary goal of a number of public Facebook pages. For example, the page "Canadians for Political Accountability, Social Justice, a Green Future," [70] which has roughly 16,000 followers, does frequently share scientific articles, but it appears to be primarily concerned with using science to highlight political problems and express opposition toward governmental figures. Their stated interests are: "Democracy, Advocacy, Politics, Environment, Future, Accountability, Transparency." This particular page appears to have been created on July 4, 2012 with the title "National Stop Harper Campaign" as an effort to challenge the current prime minister of Canada at that time, Stephen Harper. The page name was changed four days after the Canadian federal election in which Justin Trudeau defeated Harper.

Indeed, “Environmental Conservation Organization” is one of the categories of pages that content providers can apply to their page. Seven pages in our dataset identified as this category of page. The largest of these pages recorded in our dataset is the “Alt National Park Service” page [71] which has about 2.2 million followers. This page was created on January 26, 2017, six days after the inauguration of Donald Trump as the 45th president of the United States, with the stated mission of “[standing] up for the National Park Service to help protect and preserve the environment for present and future generations.” As they write in their “About” section: “We formed in response to the new administration, who has shown little mercy for the environment.” This page appears to use scientific findings primarily as a way to counter deregulation efforts it perceives as harmful to the environment, but at times also shares research that advocates for the benefits of natural conservation.

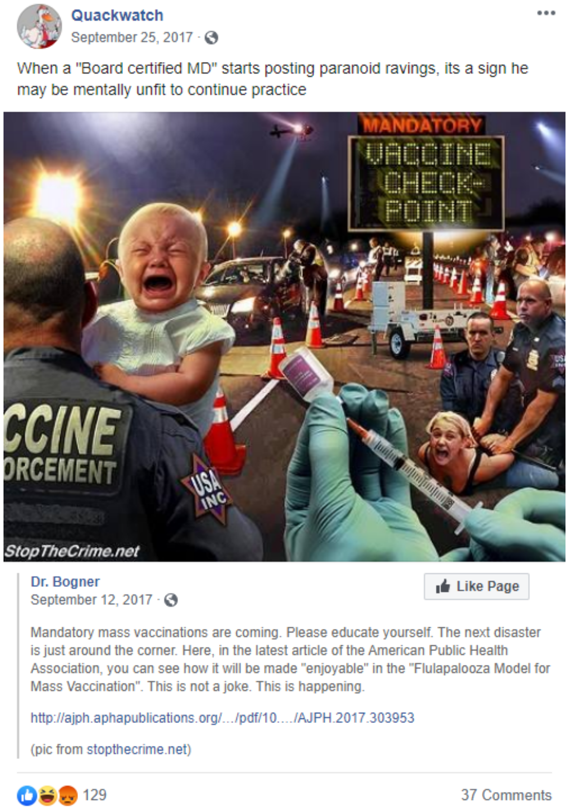
Other pages also shared content related to political issues, but with a more narrow focus on specific topics. The page “Rethink The Link” [72] is primarily concerned that Australia’s Perth Freight Link, which was announced in May 2014 and subsequently canceled due to a change in state government in March 2017, “carves through significant wetlands,” and that there may be long-term health problems caused by the construction and loss of natural resources. Their “About” section poses the question—“Why waste $1.9+ billion on the Perth Freight Link when there are better solutions?” The post in our collection was made on February 3, 2017, and began: “WE CAN WIN THIS...What it takes to win an environmental campaign.” The page has been relatively inactive in the past year, posting about once a month up to its last post in September 2019.

The “Australian Family Association” page [73], which has almost 16 thousand followers and shared an article citing a study on ethical backlashes to business marketing strategies [74], is run by a political organization that stands for "the promotion of the natural and extended family...to ensure governments assist and not hinder the natural and inherent social functions of the human family," and mostly shares content promoting the dangers of same-sex marriage and advocates against changes to marriage laws by the Australian government.

Some pages, such as the "Threatened Species Commissioner," [75] which had around 40 thousand followers and shared an article about the effect free-roaming house-cats have on local bird populations in Australia [76], are nominally maintained by governmental officials and function as a place where citizens could interact with officials, ask questions of the officials, or express their own viewpoints in response to shared content. Other pages represent individual politicians, such as "Mike Bloomberg" [77], a page that has 895 thousand followers. This page spent the most money on paid advertisements of any of the others in our set—almost $41 million between May 2018 and February 2020.

3.1.1.4 Health Related Pages

Many pages appeared to be center around health care issues. Some of these pages were operated by private enterprises. Pages such as "Personal Trainer Luca Gorgoglione," [78] which has roughly 30 thousand followers, promoted dietary- and fitness-based research. The "Lindo Bacon Community" [79] has nearly 13 thousand followers, describes the page's owner as "a scientist, acclaimed international speaker, and leading advocate for body positivity" who "is changing lives through teaching, research, writing, and transformative workshops and seminars." Most of these pages promoted personal brands or business ventures; for example, the "About" section for the "Personal Trainer Luca Gorgoglione" page describes the proprietor as offering personal fitness coaching for those near his location and "distance



Quackwatch
September 25, 2017 ·

When a "Board certified MD" starts posting paranoid ravings, its a sign he may be mentally unfit to continue practice

Dr. Bogner
September 12, 2017 ·

Mandatory mass vaccinations are coming. Please educate yourself. The next disaster is just around the corner. Here, in the latest article of the American Public Health Association, you can see how it will be made "enjoyable" in the "Flulapalooza Model for Mass Vaccination". This is not a joke. This is happening.

http://ajph.aphapublications.org/.../pdf/10..../AJPH.2017.303953

(pic from stopthecrime.net)

129 37 Comments

Figure 3.2: A post shared on “Quackwatch” on September 25, 2017.

training” for others. It enumerates “10 reasons...to train with the personal trainer Luca.” The page for “Menno Henselmans,” [80] which has about 40 thousand followers, serves a similar purpose: they offer “premium online courses and online coaching” through their page, advertising the benefits of what they call “Bayesian Bodybuilding,” which they describe as “optimizing fat loss, muscle growth and strength development based on the most reasonable interpretation of the available data.”

Other pages take a different tack, and try to draw attention to bad information on the internet for the sake of disproving it. “Quackwatch” [81] has the stated mission: “To educate about and expose health-related frauds, myths, fads, fallacies, and misconduct.” Figure 3.2 displays a post collected in our dataset which shares another post in our data from the page “Dr. Bogner”—an anti-vaccination page that advocates for alternative treatments to autism.

There is not always a clear distinction between subjects. The page “National Universal Medicare For All” [82] shares research related to health care, but its posts focus on the political and policy outcomes of this research. It has just over 85 thousand followers, and has spent just under $500 on paid advertisements for political posts. Several of their past ads have been removed because they were found to violate Facebook’s policies. The page titled “Dr. Bogner,” [83] which has about 13 thousand followers, advertises for a medical business in Brighton, Michigan that offers a novel treatment called “Hyperbaric Oxygen Therapy” as a treatment for autism. That page also posts about politics, sometimes even on issues tangentially related to health.

The stated mission of “The Commonwealth Fund” [84] is “to promote a high performing health care system that achieves better access, improved quality, and greater efficiency, particularly for society’s most vulnerable, including low-income people, the uninsured, minority Americans, young children, and elderly adults.” Just over 42 thousand people followed the page, and the post in Dataset B1 included the alarming text: “Out-of-pocket spending for a woman in need of maternity care could rise by $7,000 if maternity benefits were dropped.” This post received more “Angry” reactions than any other type (291 ”Angry,” 278 ”Likes,” 137 ”Sad,” and 24 other reactions), had 104 comments and 512 “Reshares.” The page has spent just under $138k on advertisements between May 2018 and February 2020, making it an outlier in our set. Recent paid posts advertise episodes of the organization’s podcast, raise awareness of judicial rulings regarding health care, and describe health care solutions in European countries for Americans (whom the advertisements target). The “New England Journal of Medicine” is also a very large page that focuses on sharing health-related research.

3.1.2 Purpose of Pages

3.1.2.1 Sharing New Research

Cornell University's open source research archive known as arXiv[1] moderates several public Facebook pages [85, 86, 87, 88, 89]. arXiv is devoted to sharing new research openly, and is the home of 1.66 million research articles from a variety of disciplines. Their Facebook pages include "arXiv-social," which is devoted to "bringing hard science to social media," "arXiv Sanity," a "Facebook version" of arXiv's website, and a Korean version of the same page called "arXiv SanityKR." "arXiv Sanity" follows a convention that several other pages in our set follow in which the "pinned post" at the top of their page is a description of their mission. It states that the "posts (papers) will be automatically updated hourly."

3.1.2.2 Selling Products and Soliciting Donations

Several pages appeared to sell products also. For example, the "Alt National Park Service" ran a pair of paid advertisements between August 27 and September 1, 2018, for which they spent a total of $322, advertising products sold by the organization and linked to a store on the organization's website. Some pages ran paid advertisements that were aimed at increasing donations. For example, the "Wolf Conservation Center," [90] which has just over 5 million followers, spent $450 between February 28 and March 8, 2019 to promote posts that directed viewers to a page on their website that took donations. This page was explicit about its non-for-profit status, stating specifically that the page represents "a 501(c)(3) not-for-profit in South Salem, NY."

[1] `https://arxiv.org`.

3.1.2.3 Community Standards

Out of the 260 unique public pages in our set, 25 had guidelines for user behavior—about 9.6%. Most community standards are set out in the "About" section of the page. These standards tend to be framed in a positive way that encourages people to interact politely with content and other users. Moderators of "The Conversation" [63] state that they "want the discussion to be...more illuminating than the original article." The "Scientific American" page claims that they want to "encourage open discussion (which includes disagreement)." The page called "News from Science," [91] which is affiliated with Science magazine, states that they believe "in the free exchange of ideas and [encourage] all visitors...to engage in spirited conversation." They outline that though they "expect and encourage differences in onions and perspectives, [they] demand a civil discourse."

Some pages even appeal to the right to protect the proprietary information of the page-owners or other interested parties. "Science," [92] for example, lists that posts and comments must "protect confidential, proprietary, and embargoed [American Association for the Advancement of Science] and Science information." The "American Society for Microbiology" writes that they regularly monitor and remove posts "that disclose confidential or proprietary information; use the organizations trademark for personal gain; market third party information; harass, threaten or discriminate against other users; violate laws or ASM rules [sic]."

Sometimes the community standards do not take the form of a list, and instead are presented in paragraph form in the "About" section of a page. The "Wolf Conservation Center," [90] for example has the following passage:

> It is Wolf Conservation Center's policy not to delete comments posted by the Facebook community, though we may make exceptions when those comments involve personal

> attacks, obscenity and/or ethnic slurs. Posts from community members do not necessarily represent those of the [Wolf Conservation Center].

Highlighting personal attacks and obscenities is a common feature of these standards.

Most pages that have community standards outline them in the “About” section of their page, but a small number made posts with them listed and pinned them to their wall. “Quackwatch,” for example, set as its first post a message that outlines its comment policy: “respect others, no [cursing], make your point once and only once, no promotion of, or apology for quackery,..., no politics except where it DIRECTLY relates to healthcare, no calls for violence...”

Most community standards are short and concise, but some are lengthy and go into a great amount of detail. “Skeptical Science” [62] has a section titled “COMMENTING POLICY” that is 556 words long, and includes very specific instructions about what specific types of comments and content it will delete, which specific words it tends to delete (“religion,” “conspiracy,” “alarmist,” and “denier”), and the prohibition of using all capital letters, which it describes as “shouting.” It even describes the tone of comments it would like to have on its page. “The Conversation” also has a lengthy description of its standards for behavior, including 497 words. It also includes examples of bad and good modes of interaction. One of their rules stands out as unique: “BE YOU” the page writes. “We require real names...We reserve the right to delete comments made under aliases.” It does not specify what exactly it means by an “alias,” or how it discovers accounts that violate this guideline. Most of rules of the page focus on respecting fellow users. “We’re here to talk about ideas, not the people behind them.”

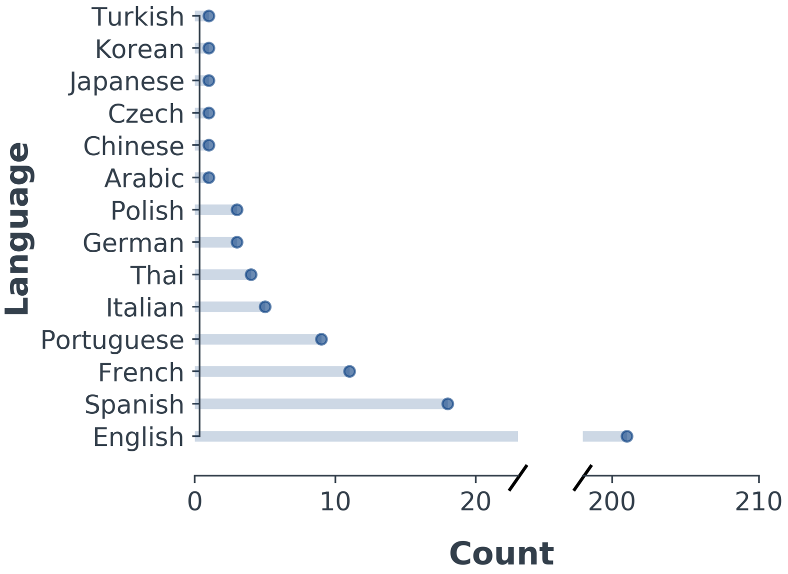


Figure 3.3: Histogram of the number of unique pages of each language in Datasets B1 and B2.

3.1.2.4 Languages

We were interested in the different languages represented in our set of public Facebook pages of science. In our set of 260 unique pages (the count that includes the combined 395 posts between Datasets B1 and B2), we found pages in 14 different languages. Figure 3.3 shows a histogram of the number of pages in each language we found in our set. The most common language was by far English, which was the primary language of 201 pages. Spanish had the second most with 18 pages, and French and Portuguese were just under that with 11 and 9 pages respectively.

On of the larger pages in our collection was a Thai page called "The Matter" [93]. It has just over a million followers, and was the source of four articles in Datasets B1 and B2.

Some pages are explicitly aimed at bringing scientific research from English into other languages to help educate specific population groups. "Lebanese Researchers" [94] is a page with almost 337 thousand followers that “aims to spread awareness and scientific thinking in the Arab community and convey the latest scientific news in an easy, simple, and enjoyable language.” They have a very broad mission, which they describe at length in their “About” section: their goals of “educating the community and transferring scientific news” into Arabic. They “believe that solving all the problems of society begins with science and education,” and hope to “make the voice of science rise above every other voice in the hope of a better future for us and our children.”

3.2 Identifying Outliers

We are interested in the relationship between the number of “Likes” a post receives, the number of special reactions, and the number of followers of the page it is shared on. Figure 3.4 shows the posts in Dataset B1 on a scatter plot. The number of “Likes” are shown on the x-axis, and the combined special reactions are shown on the y-axis. The color and size of the points show the number of followers on the page to which the post was shared. All dimensions are shown on a logarithmic scale. We are interested here in the distribution of points across the fields, and in this regard we notice that there are a number of points that lie above the main group, and one that lies significantly below. A simple ratio of special reactions to “Likes” (which we denote as Φ) and its inverse (Φ^{-1}) will give us the ability to discriminate between points that are outliers in one direction or another. The red dashed line above the main grouping of points in Figure 3.4 separates those points that fall more than 3σ (three standard deviations) above the average of Φ (μ_{Φ}). The red line below the separates the one point that falls 3σ away from the mean of Φ^{-1} ($\mu_{\Phi^{-1}}$). The point in the bottom right of the figure lies 5.34 standard deviations away from μ_{Φ}.

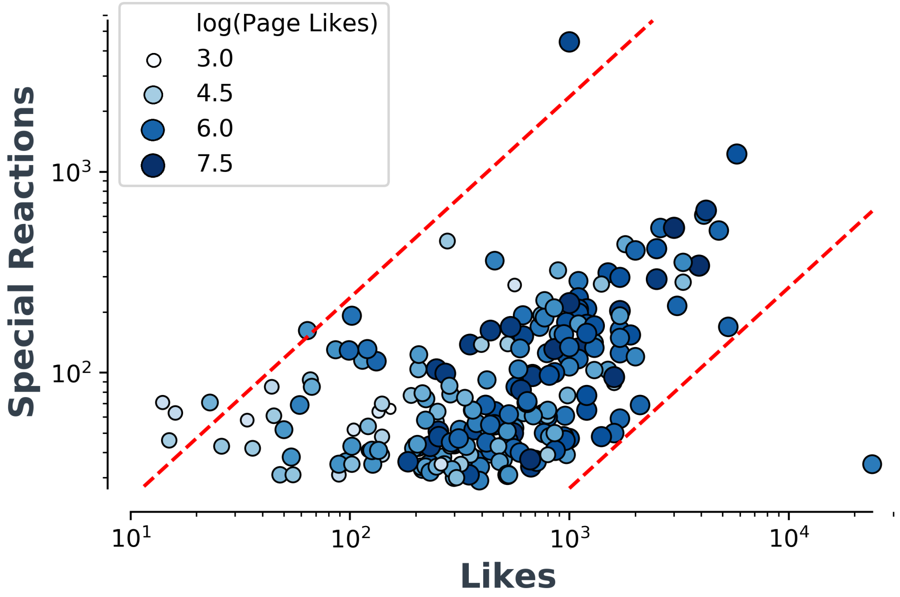


Figure 3.4: Scatter plot of posts in Dataset B1, showing “Likes” plotted against the count of the five special reactions on logarithmic scales. The size and color show the number of “Page Likes” of the public page onto which the post was made, also logarithmically scaled.

Having identified posts that lie significantly far away from the average of these ratios Φ and Φ^{-1}, it is worth investigating what content might produce these unusual responses. There are six posts above the threshold. The point at the top represents a post on the “Alt National Parks Service” page we saw earlier. The post from June 19, 2017 begins: “EPA [(Environmental Protection Agency)] ALERT: [Donald] Trump wants to eliminate EPA research program STAR” (Science to Achieve Results). It received 3.5 thousand “Anger” reactions, 1,000 “Likes,” 848 “Sad” reactions, and a lesser amount of the other special reactions. Many of the comments focus on the politics involved, especially on the reference to the U.S. president.

Another one of these outliers is the post that originally shared the picture in Figure 3.2. This pieced-together photograph is from a page called “Dr. Bogner” [83] that frequently shares anti-vaccination content that links vaccines to autism. Another post was made to a

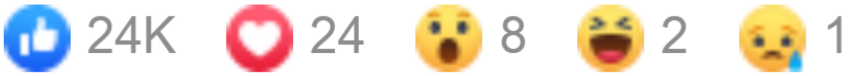


Figure 3.5: Reactions to the March 19, 2017 post on “World Bank Publications.”

page called “Exposing Feminism” [95] with 31.5 thousand followers. It shares an article [96] which it refers to as “garbage” in which a “Feminist researcher invents ‘intersectional quantum physics’ to fight ‘oppression’ of Newton [sic].”[2] One of the others was a post about the danger of corporate support for marriage equality in Australia on the page “Australian Family Association” mentioned above. Another is an alarming post on “The Conversation” with the text “A new study today has found about 15% of deaths in nursing homes are premature. And the culprit isn’t old age—they’re dying due to injury and violence.” The last one is on the page “National Universal Medicare For All,” and blames the U.S. Republican party for wanting to “tax mommies.”

The post in the lower right-hand corner of Figure 3.4 was made to a page called “World Bank Publications,” [97] which has about half a million followers. “Vietnam is at a crossroads” the post claims. “It can grow as an export platform for global value chains...or it can leverage the current wave of growth...to diversify and move up the chain into higher value-added functions.” It ends with a call for action aimed at the Vietnam government: “Success will require Vietnam’s policymakers to view the processes of development differently, and to take new realities of the global economy more fully into account.” This post was made on March 19, 2017 and received a shocking number of reactions, which are shown in Figure 3.5. It exceeded 24 thousand “Likes” but only received a small number of special reactions.

[2] The article cited in this post [96] was widely shared in similar contexts.

CHAPTER 4

MEASURING EMOTIONAL DIVERSITY

> *Diversity and independence are important because the best collective decisions are the product of disagreement and contest, not consensus or compromise.*
>
> James Surowiecki

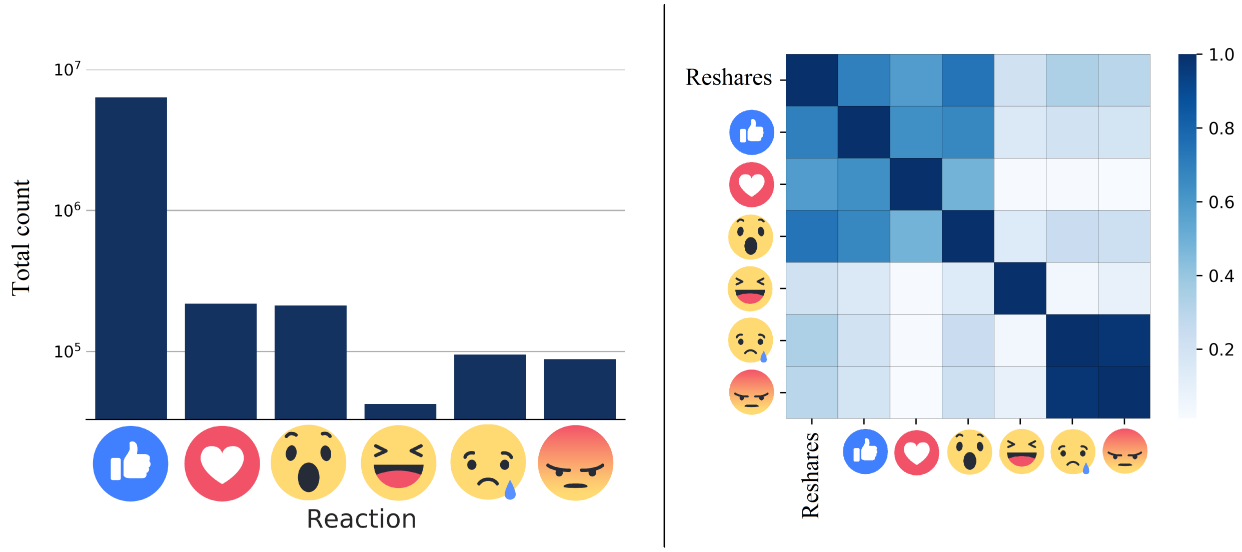
We made the claim in our introduction that advancements in scientific knowledge are often preceded by advancements in measurement—the study presented in this chapter develops from that idea. We want to find new ways to measure emotional responses of groups using new features on social-media platforms. We want to develop metrics that use click-based reactions to better understand the emotional response to social-media posts of scientific findings. Here, we are specifically interested in quantifying the ideas of disagreement, consensus, and compromise with our data. Understanding how groups respond to scientific findings is useful for a number of reasons: it can help science communicators better frame their content to address specific emotional configurations, and it can help scientists and scientific institutions more appropriately and effectively allocate resources toward maximizing positive outcomes. The influence that emotions have on the public understanding and acceptance of scientific findings has been overlooked for too long, and this study is aimed primarily at using novel features to fix this oversight.

We are primarily interested in quantifying three basic categories of emotional response: we look to capture the emotional valence, intensity, and diversity using click-based reactions. Before outlining how we transformed and used the features in Dataset A to measure these emotional categories, we will first describe some of the basic patterns and relationships we found in our data that informed the decisions we made. Finding the appropriate way to assign importance to this feature or that requires knowledge of what each one signifies and how the features interact with one another. Changes to features must be made carefully and with much deliberation, as important information can be lost or disfigured through the transformation process.

4.1 Preliminary Analysis

4.1.1 Descriptive Statistics

Click-based reactions are not evenly distributed throughout our datasets. Figure 4.1a shows the total number of each reaction type in our Dataset A. We see that there are more “Likes” than any other reaction by an order of magnitude. Among the five special reactions, “Love” and “Wow” are prevalent while negatively valenced reactions such as “Sad” and “Anger” are less common. Likewise, Table 4.1 shows descriptive statistics on the six click-based reactions and “Reshares.” We might hypothesize that Facebook users are more likely to react positively toward scientific content (or to content on the platform in general), or that positively valenced scientific content is more likely to be propagated through the platform. We might also think that these positive reactions are more common because they are more physically accessible to the user than negatively valenced reactions. While these hypotheses may well carry some weight, we can learn more through closer scrutiny of reaction usage. There are three main factors we will consider when looking at Figure 4.1a: (i) the historical timeline in which Facebook released reactions, (ii) the layout of the Facebook user interface, and (iii) the closeness in semantic meaning of the terms “Like,” “Love,” and “Wow.”



(a) (b)

Figure 4.1: Subfigure (a) shows the distribution of click-based reactions from Dataset A; (b) is a heatmap of the Spearman correlation coefficients between click-based reactions to posts and “Reshares.” Both visualize values before feature transformation.

Table 4.1: Descriptive statistics for the six click-based reactions and “Reshares.”

Reaction	**Statistics**						
	Mean	Std. Dev.	Min.	25%	50%	75%	Max.
Like	171.2	1169.1	0	13	37	107	156,613
Love	6.5	49.3	0	1	1	3	6,026
Wow	6.3	52.9	0	0	0	2	4,120
Laughter	1.3	26.9	0	0	0	0	4,162
Sad	2.8	34.7	0	0	0	0	2,778
Anger	2.6	48.6	0	0	0	0	3,978
Reshares	60.9	386.1	0	3	11	37	38,272

First, “Like” was the original reaction provided by Facebook. Between 2004 and 2016, “Like” and “Reshare” were the only click-based reactions available on the platform. By the time the five special reactions were released in 2016, users were well accustomed to using “Likes” to respond to a variety of content (for example, positively and negatively valenced posts). It follows that even with the more extensive reaction palette, users were still more likely to employ “Likes” out of habit. Second, the five special reactions are reached in the user interface by hovering over the “Like” button to open the special reaction palette. “Sad” and “Anger” are positioned on the far right side of the palette and therefore take the most effort and intention for the user to select, while the “Like,” “Love,” “Laughter,” and “Wow” reactions are grouped on the left side of the palette. The ease with which a user can click “Like” and that reaction’s spatial association with “Love” and “Wow” might account for these reactions’ prevalence in Dataset A. This argument is cogent, but can only be extended so far.

Finally, the term “Like” is more semantically related to “Love” and “Wow” (as in amazement or awe) than any other of the reactions. Further, the icon that represents a “Like” is a thumbs-up, a gesture of support and agreement that generally expresses positive sentiment; its positivity can be more closely tied to “Love” (which is represented by a heart) and “Wow” (represented by an amazed face). These associations account for at least some of the use of this feature. Studies such as Sumner et al. [98] have also shown that “Likes” are used more “faithfully” by users—a category that designates that a feature’s usage is in line with the designer’s original intent.

We should be careful, however, not to stretch this semantic correlation too far. For example, we might be surprised to find that friend’s post about a deceased relative or personal hardship has received “Likes”; clearly these responses are not meant to show that his or her friends are happy about the circumstances, but rather that they are expressing something more akin to solidarity or sympathy. This example shows that the use of the “Like” reaction

Table 4.2: Proportion of articles that received one or more of each reaction.

	Proportion
Like	0.987
Love	0.757
Wow	0.487
Laughter	0.131
Sad	0.174
Anger	0.116

Table 4.3: Proportion of articles that contain one or more of each reaction per pair of reactions.

	Like	**Love**	**Wow**	**Laughter**	**Sad**	**Anger**
Like	-	0.750	0.483	0.130	0.172	0.115
Love		-	0.310	0.102	0.097	0.067
Wow			-	0.098	0.116	0.081
Laughter				-	0.049	0.043
Sad					-	0.076
Anger						-

is not necessarily tied to its semantic meaning. Sumner et al. [98] have identified this same effect through surveying a group of 255 individuals about their click-based interactions with content on Facebook. They found that the usage of the "Like" button was less deliberative and more automatic than the usage of special reactions.

Table 4.2 shows the proportion of articles in Dataset A that contain each reaction type. Table 4.3 shows the proportion of articles in our dataset that received at least one of both reactions in each pairing, provide evidence for this effect. We can tell by comparing the proportions of articles with "Likes" and other reactions (row 1 of Table 4.3) to the values in Table 4.2 that "Likes" are paired with all other reactions almost any time the five special reactions are used. The correlations in Figure 4.1b indicate that "Likes" co-vary with "Love" and "Wow" reactions, but that does not mean that they are not paired with other reaction types as well.

Since “Love” and “Wow” share a semantic and physical closeness to the “Like” reaction, we would expect not only to see that they also are used more frequently, but also that the usage of these features can be shown to have positive correlation. Figure 4.1b displays the Spearman correlation coefficients between features. There is a high positive correlation between “Like,” “Love,” and “Wow.” Our intuition that where we see “Likes” increase, we also expect “Love” or “Wow” reactions to increase also is supported by our data. The negatively valenced reactions “Sad” and “Anger” are also highly correlated with each other. The high positive correlation between “Reshares” and “Like,” “Love,” and “Wow” reactions leads us to believe that positive content is more widely shared and reacted to on Facebook—a finding that goes against the conclusion of studies such as [46], which showed that negative emotions lead to greater interaction and dispersion across social networks.

For precisely these reasons, we should not undervalue the appearance of less common reactions such as “Anger” or “Sad.” Their appearance on an article represents more intentionality and effort on the part of a user to provide a specific response. Sumner et al. [98] have also found that the five special reactions “were perceived as more deliberate and less automatic communicative behaviors than Likes” by respondents to their survey. Because positive reactions are the expected mode of response, and because negative reactions take more effort for user to apply, we decided to weight the different kinds of reactions by the inverse proportion of our expectation of seeing a reaction of that type.

4.2 Feature Transformation

Articles in our datasets are not equally reacted to. A share of an article on one page may result in thousands of responses, while sharing that same article onto another page may result in no reactions at all. The number of reactions may be a signal that a post or article evokes a strong response from users, but it is difficult to account for the context, such as the number of people who saw the post. How many users follow public pages varies from page to page and so the visibility of a post on each page will also vary. It is also difficult to account for Facebook's algorithms that propagate posts into users' news feeds.

Our method of weighting the click-based reactions is based on the probability with which we would expect to find them on any given article. Weights were determined using a method that is related to Term Frequency-Inverse Document Frequency (TF-IDF)—a method of setting the relative importance of terms within a body of documents that is used in many information-retrieval and recommendation systems today. We refer to the weighting procedure we used as Reaction Frequency-Inverse Document Frequency (RF-IDF).

We changed the raw reaction counts for each article into a proportion of all the click-based reactions the article received. For example, if an article received 6 "Likes," 1 "Love," and 2 "Wow" reactions, we would transform those values into 6/(6 + 1 + 2) "Likes," 1/9 "Love," and 2/9 "Wow" reactions. We then logarithmically scaled these values. The result of this transformation gave us the *Reaction Frequencies* (RF_{d_r}) (Equation 4.1), where d_r is the count that a given document d received for a certain reaction r, and R is the list of all six click-based reactions.

$$\mathrm{RF}_{d_r} = \log \frac{d_r}{\sum_{r \in R} d_r} \tag{4.1}$$

Next, we needed to reward the rare reactions and penalize the common reactions and to determine the appropriate weights for each reaction type. To do this, we found the probability that a given kind of reaction will be applied to a random article, which is found by taking the number of articles in our dataset that had that reaction type and dividing that number by the number of articles in our dataset. The Inverse Document Frequency (IDF) is the natural logarithm of the inversion of this probability. This value gave us the IDF for each reaction type (Equation 4.2), where $|D|$ is the number of articles in the dataset and $|D_r|$ is the number of articles in the dataset that received certain reaction r.

$$\mathrm{IDF}_r = \log \frac{|D|}{|D_r|} \tag{4.2}$$

Finally we computed the RF-IDF by multiplying the logarithmically scaled proportion of each reaction type for each article by the IDF for that kind of reaction, as shown in Equation 4.3.

$$\text{RF-IDF}_{d_r} = \mathrm{RF}_{d_r} \cdot \mathrm{IDF}_r \tag{4.3}$$

4.3 Metrics of Emotional Diversity and Intensity

With our transformed click-based reactions, we developed metrics to measure the valence, intensity, and diversity of user responses to the articles in our dataset. Valence is the most simple of the three metrics. Since it represents the positive or negative direction toward which a response tends, we had to determine the signals of positive and negative emotions coded in reactions. “Love” and “Anger” are relatively straight forward, having positive and negative valence respectively. The “Sad” reaction could be used to express sympathy or solidarity with a person who has undergone some difficulty, but in the set of posts we are studying it is not likely that users are consistently sharing personal experiences in their posts or research that could inspire this type of sympathetic reaction. Figure 4.1b shows a very high correlation coefficient between “Anger” and “Sad”—indeed this pair of features have the highest correlation of all feature pairs. This provides evidence for the idea that these two reactions share a common valence.

To determine the valence of each article, we thus checked to see whether the value of its “Love” reactions was greater than that of its “Sad” and “Anger” reactions, as shown in Equation 4.4.

$$d_{valence} = \begin{cases} -1, & \text{if } (d_{sad} + d_{anger}) > d_{love} \\ +1, & \text{else} \end{cases} \tag{4.4}$$

Next we computed the intensity of the response for each article in our dataset. In conceiving our metric, we started with the observation that when providing a click-based reaction to any post on Facebook, each user gets exactly one reaction they can provide, not including “Reshares” which can be selected by a user who has “Liked” or provided one of the five special reactions. By choosing to select one of the special reactions over the default “Like” a user is demonstrating a desire for a more specific response. We interpret this intention and effort as a sign of a stronger emotional reaction. With this observation in mind, we considered intensity as the ratio of the five special reactions to the sum of the six click-based reactions. We began by summing all the reactions for a given document d, as shown in Equation 4.5.

$$d_{total\_reacts} = d_{like} + d_{love} + d_{wow} + d_{laughter} + d_{sad} + d_{anger} \tag{4.5}$$

We then summed up the five special reactions and divided them by the total click-based reactions, as shown in Equation 4.6. This bounded our intensity metric between [0,1], where 0 represents an article that received “Like” reactions but nothing else, and 1 being the score of an article that received only the special reactions; bounding our metric in this way facilitated comparison between articles. We designed our intensity metric to be sensitive to posts that receive a low reaction count. It was important to us that it be able to identify posts that receive strong signals of emotional intensity without relying on the sheer quantity of reactions.

$$d_{intensity} = \frac{d_{love} + d_{wow} + d_{laughter} + d_{sad} + d_{anger}}{d_{total\_reacts}} \tag{4.6}$$

Finally, we developed a metric to measure the diversity of user responses to each article. Diversity measures how many of the different special reaction types are present for a given article as well as how evenly distributed those reactions are. We disregarded the "Like" reaction for our diversity measure. As discussed above, this reaction is quite flexible in use and relies on context for meaning. We use Jensen-Shannon distance (JSD), which uses entropy and Kullback-Leibler divergence to measure the difference between two probability distributions, as the basis for our metric. JSD is found by taking the square root of the Jensen-Shannon divergence score. We preferred JS distance over JS divergence because the former is a true metric of distance and has been shown to satisfy triangle inequality. This latter property improves our ability to compare the results for multiple articles. We preferred JSD over Kullback-Leibler divergence because the latter measure has no upper bound, making comparison between different observations of the measurement difficult.

We took as a given that the highest diversity of reaction types that could be observed is a uniform spread of types where each reaction is present in equal proportion, and recorded the JSD between this uniform and our observed distribution. JSD is a (0,1) bounded value, where (in our case) 0 indicates that the observed distribution is uniform and 1 shows that an observed distribution varies greatly from uniform (i.e., has only one type of special reaction).

$$JSD(P \parallel Q) = \sqrt{\frac{1}{2}KLD(P \parallel M) + \frac{1}{2}KLD(Q \parallel M)} \tag{4.7}$$

where P and Q are two distributions, $M = \frac{1}{2}(P+Q)$, and $KLD(P \parallel Q)$ is Kullback-Leibler divergence. Since a uniform distribution, which represents the most diversity an article can receive, will produce 0, we took the complement value:

$$d_{diversity} = 1 - JSD(\theta_d \parallel \mathcal{U}\{0,1\}) \tag{4.8}$$

where θ_d is the distribution of the five special reactions to a given article d and $\mathcal{U}\{0,1\}$ is the discrete uniform distribution from [0,1].

Our metrics are designed to be combined to allow further comparison across articles. For example, diversity evaluates how many different reactions are present on a given article, but not how many reactions are present, or the proportion of the five special reactions to all click-based reactions. By multiplying the diversity by the intensity as displayed in Equation 4.9, we are able to identify articles that received both an intense response, as well as a response that has many different emotions present.

$$d_{divint\_index} = d_{diversity} \cdot d_{intensity} \tag{4.9}$$

We can also combine valence and intensity scores to produce a polarity score, which reports both strength and direction of a response (Equation 4.10).

$$d_{polarity} = d_{valence} \cdot d_{intensity} \tag{4.10}$$

4.4 Experimental Design

4.4.1 Training a Topic Model

Our interest lies beyond how individuals are responding to specific scientific articles: we want to use our data to gain a better understanding of aggregate emotional responses to areas of science. To do so, we wanted to group papers in a logical manner as possible. We initially considered using the "Scopus subject" tags that were associated with each article in the Altmetric dataset, but a great number of articles were missing this feature. Furthermore, users who respond to a post do not necessarily click the link to open the article and read the content before reacting, as shown in [27]. It follows that users are reacting to the content they directly encounter: the Facebook post. As such, we wanted to find ways to maintain as much of the granular detail of the post texts as possible. Following this line of reasoning, we trained our topic model on the texts included with the article shares of each article.

We took the following steps in building our LDA topic model. We began by combining each text that accompanied shares of a given article. Each article can be shared many times, therefore our "texts" were of widely varying length. LDA topic models are especially powerful tools for handling data of this kind [24]. These posts can also be in a variety of languages, so we kept only texts that were in English. We cleaned the texts, removing hyperlinks, punctuation, email addresses, and hashtags. We then created bi-grams and tri-grams (common groupings of two and three words, respectively) from our tokenized texts. We removed stop words such as "the" and "and," as well as lemmatized each word, which involves removing inflected endings to transform each word into its dictionary form (e.g., making plural nouns singular and changing verbs into their infinitive form). Finally, we removed words that appeared in less than fifteen documents and those that were in more than half the documents. The result was the corpus we used to train our model.

We used the LDA model from Python's Gensim library [99]. Selecting the number of topics t that a model should identify can be time-consuming. The number may change depending on the goal a researcher has for a model or by the form of their data. Our goal was interpretability: we wanted to be able to easily grasp what a given topic is about, as well as to be able to distinguish one topic from another with relative ease. We also wanted the number of topics to reflect the number of fields we would expect to see in our dataset. There were about thirty different Scopus subject labels (e.g., physics, biochemistry, computer science) applied to the articles in our dataset. We reasoned that setting t somewhere between 15 and 40 would give an appropriate representation of the fields we expect to see in our set.

We trained seven LDA models, each with a different value of t across the range: $15 \leq t \leq 50$. We compared our models based on their *topic coherence scores* (CS), which is a measurement that tests the degree of semantic similarity between the most representative words in each topic. A higher CS score usually indicates that a model has better identified distinctive topics. The highest coherence score in our set of models was $CS(t = 20) = 0.515$. We then used our model with $t = 20$ transform the post texts into a distribution of topics. LDA returns the probability that each topic is present for a given document, so we selected the topic that had the highest probability of being present to represent each article. For example, if document x had the topic distribution of $[t_1 = 0.7, t_2 = 0.3]$, we would label it as representing topic t_1. Table 4.4 shows: (i) the twenty discovered topics, (ii) the number of articles in which each topic was the most representative, and (iii) the ten most representative words in each topic.

Table 4.4: The twenty topics discovered by our LDA model with the number of articles where that topic was the most prominent, as well as the top ten words representing each topic.

Topic no.	***Article count***	***Top ten words***
1	1,506	australia, australian, public, year, government, national, country, policy, would, migraine
2	4,045	study, increase, show, may, high, find, level, result, low, effect
3	7,929	not, people, time, say, make, take, do, may, get, many
4	613	health, care, need, school, learn, support, practice, student, provide, help
5	378	vitamin, diet, exercise, risk, health, muscle, protein, intake, eat, body
6	3	state, power, law, grant, core, centre, fire, press, panel, legal
7	852	species, population, human, ancient, new, region, area, find, reveal, modern
8	125	vaccine, disease, infection, immune system, bacteria, cause, gut, protect, virus, microbiome
9	168	planet, light, space, earth, sun, mass, scientist, hot, star, observation
10	98	glyphosate, food, exposure, plant, animal, environmental, soil, crop, chemical, organic
11	64	disorder, autism, mental, depression, stress, cognitive, physical, self, behavior, social
12	8	death, die, skin, florida, emergency, sea, travel, page, kill, cat
13	105	use, alcohol, cannabis, health, drug, risk, product, opioid, fda, harm
14	318	ice, water, climate change, energy, global, climate, loss, change, air, growth
15	31	child, age, old, young, adult, injury, infant, parent, family, year
16	17	woman, man, mother, baby, pregnancy, risk, birth, female, male, girl
17	931	use, system, method, technology, datum, model, device, network, process, base
18	1,660	patient, pain, treatment, risk, cancer, disease, dose, medical, chronic, medicine
19	4,417	new, research, read, article, science, publish, paper, journal, scientist, nature
20	2,157	brain, cell, human, dna, function, neuron, mouse, protein, gene, genetic

4.4.2 Kolmogorov-Smirnov Test

We were interested in testing whether the observed values of our metrics for a subset of articles belonging to a given topic significantly diverge from the values among the rest of the articles. To accomplish this, we used the two-sample Kolmogorov-Smirnov test (KS test). The KS test is a non-parametric statistical test of two continuous, one-dimensional probability distributions. It makes no assumptions about the normality of either distribution, and tests whether the two distributions were sampled from the same populations (or populations with identical distributions). The KS test gives two values: (1) the KS statistic, which is derived from the largest distance (known as the *supremum*) between the cumulative distributions of the two distributions, and (2) a p-value that indicates the significance of the observed KS statistic. The question this p-value answers can be stated as follows: if we assume that the two samples come from the same population, what is the likelihood of observing a given distance between these distributions? A small p-value ($p < .05$) indicates that the probability of seeing such a difference from samples out of the same population is low, and that we can reject the null hypothesis—which in this case indicates that the two samples come from the same population. We formulate our hypotheses in the following way:

$$H_0 : P = Q$$

$$H_a : P \neq Q$$

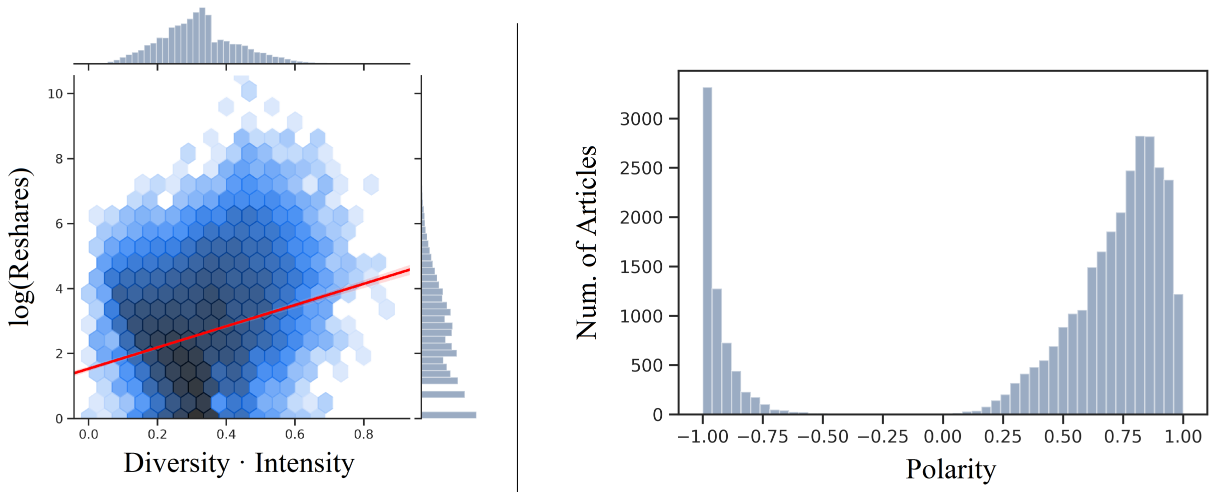
where P and Q are the two samples used as input for the KS test, H_0 is the null hypothesis, in which the two samples come from the same population, and H_a represents the case in which we reject the null hypothesis.

We set our test up as follows. (i) We chose to perform our tests with two of our metrics: diversity and polarity. (ii) We selected several topics that we hypothesized would have distributions of these metrics that departed significantly from the rest of the articles. We chose topics 1 (government), 8 (vaccines), 16 (gender), and 20 (genetics) to test for significantly diversity scores, and topics 8 (vaccines), 10 (agriculture/environmental science), 13 (drug and alcohol), and 14 (climate change) for polarity scores. (iii) We set the significance level needed to reject the null hypothesis at $\alpha = 0.05$. The KS test produces a two-tailed p-value, therefore we can reject H_0 only if $p < 0.025$ or $p > 0.975$, and fail to reject it otherwise. (iv) For each test, we partitioned our data into two groups: the articles representing the given topic and all other articles. (v) We then performed the KS test with the distributions of each metric on these given samples.

4.5 Experimental Results

Figure 4.2a displays the distribution of articles along two feature axes: the divint index on the x-axis and the logarithmically scaled "Reshare" count on the y-axis. We see that the majority of articles fall within the range of 0.3–0.4 for divint index score ($\mu = 0.322$, *med.* $= 0.313$, and $SD = 0.117$) and relatively few "Reshares" (statistics of this feature are shown in Table 4.1). As the divint index of an article increases, the likelihood of it being "Reshared" also increases slightly. This relationship is shown with a regression line plotted in red through the figure.

The distribution of the polarity scores is shown in Figure 4.2b. Though there are more articles that received a positive polarity score, the negative polarity scores are clustered closer to the extreme values. A majority of the articles (80%) received positive valence, but the most papers in any range are clustered around -1. There are around 3,250 articles that received



(a)

(b)

Figure 4.2: Subfigure (a) is a hexplot showing the distribution of articles in our dataset with regard to the divint index and the logarithmically scaled “Reshare” count; a regression line is plotted, showing a positive correlation between the two features; (b) shows the distribution of articles by polarity ($valence \cdot intensity$).

a polarity score of -1, compared to only 1,250 that received a polarity score of 1. There is a significant drop-off of articles in the range [-0.6, 0]. This behavior should not surprise us: we saw in Figure 4.1 that “Likes” were not correlated with “Sad” or “Anger” reactions, but they were highly correlated with “Love” reactions. Increased presence of “Love” reactions tend to occur with increase “Like” reactions, necessarily lowering the intensity score of these cases. On the other hand, increased “Sad” or “Anger” reactions do not tend accompany increased “Like” reactions, pushing the intensity score higher.

Our finding that Facebook users generally react positively to research is in accord with other studies that find that people are more likely to share positive than negative content [100]; but our results also indicate that people tend to react with stronger intensity when they have a negative reaction, as seen by the relative lack of articles in the moderately negative range (-0.6, 0] and the large number of papers with a score of -1. Besley and Nisbet [101] find that scientists view the public as “emotional” (as opposed to “rational”) and

Table 4.5: The results of our two-sample KS tests performed for diversity and polarity scores of four topics each.

Topic		**Test results**	
Topic No.	Keywords	KS statistic	p-value
	Diversity		
1	government	0.024	0.4
8	vaccines	0.077	0.429
16	gender	0.107	0.984
20	genetics	0.063	0.000
	Polarity		
8	vaccines	0.068	0.596
10	agr./env. science	0.151	0.021
13	drugs & alcohol	0.085	0.421
14	climate change	0.068	0.106

"fear-prone" when it comes to accepting scientific findings. The prominence of positive emotional reactions expressed toward the articles in our dataset suggests that the "emotional" response to science is on average supportive. Because negative reactions or criticism are more salient to a given person [13, 18], it is plausible that these negative reactions stand out more for scientists—especially in cases in which they are more intense. It may be that these negatively valenced but intense reactions disproportionately shape scientists' impressions of the public.

4.5.1 KS Test Results

The results of the KS tests are presented in Table 4.5. For diversity scores, topics 16 and 20 showed significant deviations from the rest of the dataset. For polarity scores, only topic 10 showed a deviation significant enough to reject the null hypothesis. The KS test is known to be more likely to result in a failure to reject the null hypothesis when testing with samples of relatively smaller sizes. Topic 16 (gender) only had 17 articles representing the sample, and so we were especially surprised to find such a significant result showing for a sample size so small.

We were also surprised by cases in which we failed to reject the null hypothesis—especially for either metric with topic 8 (vaccines). Negative emotional reactions to research about vaccines are widely visible and talked about on social-media platforms, but our data indicates that negative reactions on this subject do not significantly deviate from how users respond to research as a whole. It may be the case that because we are tracking scientific articles through Altmetric's database, we are only seeing the responses to vaccine research by those who already accept the scientific findings in that field. To find dissenting opinions we would have to look outside the domain of scientific research.

The distribution of reactions we saw in Figure 4.1a may be the result of our choice of domain for this study. If we were to study, for example, articles in major news outlets it is possible that we would find a higher representation of negative emotions. In targeting shares of scholarly research, we have chosen a domain that is generally considered to be emotionally neutral—though it is not entirely without controversial topics. But even looking at popular news sources, we hypothesize that negative reactions would flag behind positive reactions for the very same reasons (their high correlation with "Likes" and the marginal amount of extra work it takes for users to select them).

4.5.2 Final Analysis

Much social-media research is predictive in nature. The models developed by people working in this field often use features from altmetrics to predict what research outcomes will be important, who the rising stars in a field are [102], or to discover surprising articles that may have otherwise been overlooked [103]. The approach we presented in this chapter for feature transformation and generation could also be used in predictive tasks. Predictive models also rely on in-depth knowledge of the data used for training and testing; the analysis we present here could thus stand as the basis for other researchers who are interested in predicting outcomes of science using social-media data.

We used our metrics to identify aggregate emotional reactions to fields of science, rather than to individual articles, but our approach could be used for individual posts as well. Content managers or platforms might want to find negative or controversial content quickly and efficiently in order to improve user experiences. It could also help scientists and researchers better understand the role they play in shaping the emotional dynamics in broader society. Our metrics provide a way to identify this type of material that is sensitive to sparse reaction profiles so that appropriate responses can be made quickly and efficiently when necessary.

CHAPTER 5

MODELING EMOTIONS TOWARD TOPICS IN SCIENCE

A cocktail not shaken enough soon rots.

Heraclitus

Up to this point we have explored high- and low-reaction posts and numerous features of the pages onto which those posts were made; we have developed and demonstrated the value of several new metrics that we derived from our exploration of click-based reactions; we have used our metrics to understand how users are responding as a group toward topics of science; but we have not approached the question of how respondents feel in general toward topics of science. Do posts about animals make Facebook users happy, angry, or sad? Do they have strong feelings toward articles about climate change or vaccines? Do posts about the universe make them feel a sense of awe? What topics appear most frequently in our data? These are fundamental questions that we are interested in using statistical methods and machine learning models to answer.

How can we answer these questions with our data? Changing the formulation of our questions is a step in the right direction. We can pose our motivating question as follows: how likely is a post of a given topic in science to produce a specific type of reaction? Answering this question is the purpose of the experiment conducted in this chapter. To do so, we will make use of our large Dataset A to train an LDA topic model that achieves more specific topics than we did before; we will showcase a novel approach to feature transformation that relies on principles of compositional data analysis that have yet to have a wide-spread impact on the field of machine learning; and finally we will train logistic regression models that learn average emotional responses expressed as click-based reactions to topics in science. We present the results of our experiment in several tables that give a great amount of detail about the models and the statistical patterns they discovered.

5.1 Experimental Methods

5.1.1 Logistic Regression

We can describe our working problem in this chapter in the following way: we want to find a function that can tell us the likelihood that a specific text will produce a given type of response. We can represent this function as $p(Y = y \mid X = x) = \theta$, where θ is a conditional probability distribution, Y is a response feature in which $Y \in \{like, love, wow, laughter, sad, anger\}$, and X is a feature vector that represents the texts included in our Facebook posts.[1] Logistic regression provides a simple and elegant way to learn such a function. Logistic regression is related to linear regression, with the primary difference being that in the former we are trying to predict a categorical variable while in the latter case, we are predicting a continuous variable. The simplest form of this model applies to cases in which we have a binary response. For example, we may want to predict given a student's academic record whether he or she will be admitted into a top graduate program. To predict this categorical variable (the answer is either *yes* or *no*), we use the logistic function:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{5.1}$$

where β_0 is the intercept term and β_1 represents the coefficient for covariate X. This function gives as output a value between 0 and 1 for all values of X. This value represents $p(Y = 1 \mid$

[1] Much of the overview of logistic regression presented in this section comes from Gareth et al. [104].

$X = x$). In our example of a prospective graduate student, this is the probability that our student will be admitted.

To estimate the regression coefficients (β) of this model, we use a method called *maximum likelihood estimation* (MLE) [104]. MLE uses an iterative process that begins with a rough estimate of the values and corrects it until the outcome is not improved by further iteration, which is known as the model's *convergence* on the optimal solution. With some algebraic manipulation we can rework Equation 5.1 into the following form:

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X} \tag{5.2}$$

Taking the logarithm of our function gives us:

$$log(\frac{p(X)}{1-p(X)}) = \beta_0 + \beta_1 X \tag{5.3}$$

The left side of Equation 5.3 is what is known as the *log odds*, or *logit* function. This function maps probability values from $(0, 1)$ to $(-\infty, +\infty)$. Figure 5.1 shows the shape of the logit function. It shows how we can translate values that vary from negative to positive infinity (as the log odds do) into probabilities bounded between 0 and 1. The x-intercept falls at the point (0, 0.5), which represents the point where there is even probability that the response is of both categories, or $log(x) - log(1 - x) = 0$ and $x = 1 - x = 0.5$, given $0 < x < 1$.

5.1.1.1 Interpreting Logit Models

Logistic regression is often used because it provides us with a good deal of information about how the model determines the probability of a certain response. As with linear regression, the estimated coefficients $\hat{\beta}$ tell us how much an increase of one unit in the independent variable will change the response variable. In logistic regression, however, the interpretation

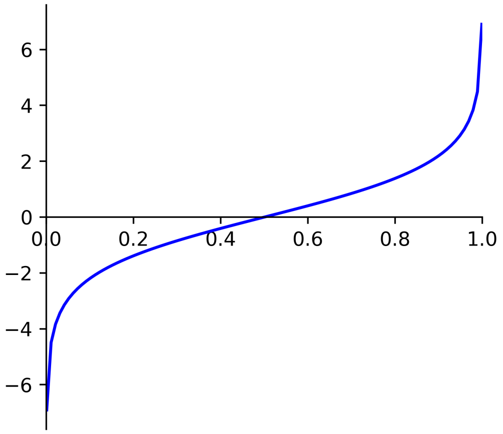


Figure 5.1: A plot of the log odds function $f(x) = log(\frac{x}{1-x})$ in the domain $(0, 1)$.

of these coefficients is slightly different: they tell us how much the model estimates a 1-unit change in the independent variable will alter the log odds. It is important to keep in mind that $p(X)$ in Equations 5.1, 5.2, and 5.3 represents the probability of success in our successive trials. What a "success" means is determined by the way we specify our categorical responses. In simple logistic regression, a binary response is typically coded as 1 for "Success" and 0 for "Failure," and therefore $p(X = 1) = p(x)$ and $p(X = 0) = 1 - p(X = 1)$. Our "Failure" class is known as the *reference* (or sometimes *referent* class) and is represented by the denominator in Equation 5.2. The logit model will not specify the probability of failure, but we can easily take the complement of $p(X)$ to discover this value. Since the logit model's estimated coefficients do not measure $p(X)$ directly but rather the log odds, we can use work backward from Equation 5.3 to 5.2 to find $p(x)$.

We can also compute confidence intervals with logistic regression, and use these to perform hypothesis tests to measure the strength of the connection between each independent variable and the response given the size of our sample we used to test this relationship and that the samples are randomly drawn from the underlying population. These tests give us a p-value, which tells us what the chances are that there is no relationship between and independent variable and the response.

5.1.1.2 Logistic Regression with Multiple Response Categories

Multinomial logistic regression extends the concept of logistic regression for problems where the response variable can take the form of more than two possible discrete outcomes. This is a generalization of simple logistic regression, but it provides us with the same ability to make inferences that we have with the simple model. The interpretation of the coefficients is slightly more complicated. Similarly to simple logistic regression, we take one of the response classes as a reference and use MLE to estimate the log odds that each of the other response classes will be selected over the reference class. As we did with our simple logistic model, we can work backward from the coefficients to produce probabilities for each of our responses given the presence of a given feature vector $\mathbf{x}$, an idea that we will develop later in this chapter.

5.1.1.3 Why Logistic Regression?

It is worth asking: why have we chosen logistic regression models over other models? We might have alternatively preferred to use a naïve Bayes classifier; or we may selected to use a more complicated and powerful deep-learning model such as a convolutional neural network. In the latter case, we have opted in favor of simpler models. Our goal is not to achieve higher predictive accuracy at the cost of interpretability—rather we are interested in doing the exact opposite. Our aim is to explain the connection between the presence of certain topics in science and certain emotional reactions. While deep-learning models might provide a better predictive accuracy for estimating which reactions a user might apply in a given scenario, we choose to sacrifice this accuracy in the name of better explanatory power from our model.

Naïve Bayes models are also relatively simple—so why have we preferred logistic models over them? Ng and Jordan [105] show that discriminative models such as logistic regression have a lower asymptotic bound for their error rate than generative models such as naïve Bayes do, though they take more data to reach that lower bound than naïve Bayes does. We are fortunate to have a very large dataset approaching 500 thousand observations. By using logistic regression, we can leverage the size of our dataset to achieve better results and gain more information about the phenomenon we wish to study than we could with generative models such as naïve Bayes.

5.1.2 New Perspective of Our Data and Model Assumptions

To answer the questions we have set up in this chapter, it will be useful to look at our data from a new perspective. Up until now, we have tended to view the reactions an article or post received as a set of counts. We have also looked at the proportion of reaction types to posts or articles. But another way to see these reactions is as categorical responses. We can re-imagine our data as an experiment where a group of participants were shown a text and asked to respond with one of six reaction types that best described how the text made them feel. This may seem like an unsurprising way to look at our data since it is closer to the way we individually experience social media, but it will be helpful to us to take this position here.

It will also prove useful to change our perspective of the posts. We have tended to focus on the unique aspects of each post, narrowing in on ways it might provoke certain aggregate responses. We can instead think of the posts as representing mixtures of topics in science. We can then see the reactions as responses to specific combinations of topics, and use them to evaluate what people's reactions are to the underlying topics.

Logistic regression has several assumptions about the data used as input that we should address. First, it assumes that the observations are independent of each other. This means that if we see a "Love" reaction, the next reaction type we see is not influenced by the first. This assumption is worth pondering in our case. Are the reactions in our dataset independent of each other? On the scope of a single post, it makes sense that this assumption will not hold. A user can see the responses that have already been applied to a post, and this information may affect the way they choose to respond. Measuring the strength of this conditional dependence would be an interesting goal, but is outside the scope of this study. On the scope of our entire dataset, however, this assumption of independence does seem to hold true. We are looking at shares across thousands of public Facebook pages, and the reactions on one page have no bearing on the reactions on another. The scale of data we are working with justifies assuming independence, but we should be careful to ensure that we do not reduce our data to a small enough sample where this assumption would no longer hold.

A second assumption of logistic regression is that there is little or no multicollinearity (or correlation) between the independent variables. Multicollinearity, or *data singularity*, occurs when a row or column of a matrix can be expressed as a linear combination of other rows or columns. This is clearly a problem in our case, as the topic distribution vectors we wish to use as representations of our documents are proportional and thus at least one column can be expressed as a linear combination of the others (i.e., knowledge of two out of three possible topic values provides you with enough information to always find the third, missing value). This is an important problem that we will have to address before using our data as we have proposed. It is enough to introduce the problem here, but we will return to it in greater detail shortly (and in even greater detail in Appendix A).

5.1.3 A Different Approach to Topic Modeling: Seeding Topics

In the study presented in Chapter 4, we used an LDA topic model to group the documents in our dataset that received the highest probability of belonging to the same topic. This approach is common in NLP research that uses LDA, but it does not make full use of the probabilistic nature of this model architecture.

The concept of LDA migrated to NLP from bioinformatics research around the turn of the twenty-first century, making a splash after the publication of Ng et al. [24] and is still used today over other types of topic models (e.g., *latent semantic analysis*—or LSA—topic models) because it represents documents as an infinite mixture of underlying topics. The motivating idea behind LDA was to view the generating properties behind a corpus of documents as probabilistic rather than deterministic. Earlier topic models had used word frequencies across documents to determine important words; but the algorithms behind these models were not as dynamic, scalable, or flexible as researchers had wanted. As the pre-Socratic Greek philosopher Heraclitus wrote: “a cocktail not shaken enough soon rots.” The idea behind LDA is that we need to use probability to “shake up” our topic mixtures so they do not stagnate, and instead generalize better and more quickly. McAuliffe and Blei [106] show that the probabilistic nature of the LDA algorithm is precisely *why* it was developed, and why it should generally be preferred over other topic model architectures. Topics cannot be observed directly and fixed, but represent the generative probability space that produces some words in a document and not others. Topics are the tectonic structures underlying a corpus of documents that produce specific word groupings, and LDA uses Bayesian statistical processes to estimate the likelihood that a given topic would produce the specific word combinations of a given document.

LDA represents topics as a probability distribution over a set of words, where some words are more likely to appear together than others, and documents as mixtures of topics. In using LDA as we did in Chapter 4, we are not making full use of this latter property of the model. In classifying a document by the topic that is most likely to represent it, we lose information that could help us to gain a better understanding of our documents and the emotional responses that Facebook users applied to them. For example, say we have two documents d_1 and d_2. Imagine that d_1 receives a 20% chance of belonging to topic 1 and a 19.99% chance of belonging to topic 2, and that d_2 receives a 70% chance of belonging to topic 1 and only a 5% chance of belonging to topic 2. With our "winner-takes-all" method of assigning topic 1 to both documents, we ignore the fact that while the majority of words in d_2 are likely produced by topic 1, d_1 consists of almost equal parts words from topic 1 and 2. By treating these two documents as belonging to the same class, we lose information about their unique properties. A bigger problem we may have is that treating a document as "belonging" to a given topic completely obscures the probabilistic nature of LDA.

We want to change the way we utilize LDA in two important ways. First, we want to take advantage of the full topic distributions that the model applies to a given document. Doing so will improve our ability to assess the response to specific topics in science. We will achieve this by letting the documents be represented by the unique topic distributions produced by our LDA model. Second, we want to increase the interpretability of the topics we discover by tuning some of the parameters of our model. LDA models allow us to use our prior knowledge of the training data to influence how the model identifies word groupings that are associated together into topics.

5.1.3.1 Inner-workings of LDA

Understanding how we accomplish this second goal requires a brief explanation of the way that the LDA algorithm works. LDA topic models estimate several distributions through training: it learns (1) $P(w_i \in t_j)$, or η (eta), the probability that each word (w_i) in a corpus belongs to each topic (t_j), where $\eta \in \mathbb{R}^{n \times k}$, n represents the number of topics we are looking for, and k represents the number of unique words in our corpus; and (2) $P(d_i \in t_j)$, or the probability θ (theta) that each document (d_i) in our set belongs to each topic (t_j), where $\theta \in \mathbb{R}^{n \times m}$, n is the number of topics, and m is the number of documents in our set. These two sets of probability distributions are represented by two matrices η and θ, and are what is "learned" by the LDA training process.

There are different methods by which η and θ are initialized before training, but most implementations allow us to specify prior distributions. We may have some prior knowledge that a specific group of words belong together inside a topic, and so we can increase the probability that these words will be associated with this topic and decrease the probability that they might belong to other topics. By doing this, we can increase the likelihood that certain word groups cohere into a topic. Specifying priors for these distributions can be useful, but by no means guarantees that our desired groups cohere. LDA may determine that the associative patterns in our documents do not support our prior belief, in which case the training process will reduce the coherence of our hand-picked word groupings and associate the words with other topics. Using prior belief about the topics in a corpus is known as *seeded* [107] or *guided* [108] LDA has resulted in significant improvement to the coherence and interpretability of topics. Using topic models in this way is what is known as "semi-supervised machine learning" as it incorporates elements of both supervised and unsupervised learning [109].

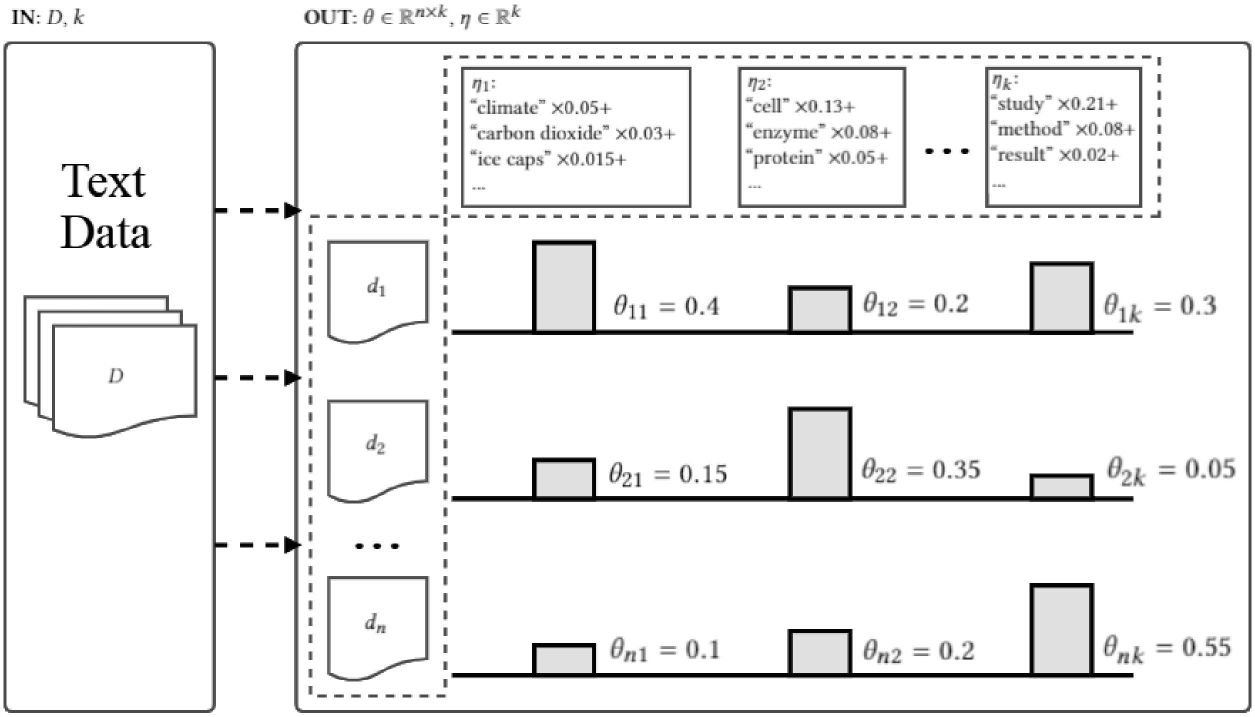


Figure 5.2: Diagram showing the basic input and output structure of an LDA topic model.

Figure 5.2 presents a basic diagram of an LDA topic model. On the left, we see the input to the model: a set of documents and number of topics k. On the right side, we see the output of LDA: θ, a distribution of topics per document, and η, a distribution of words per topic. For each of our documents $d_i \in D$, we show a hypothetical distribution of several topics. These values show the amount of each document that is predicted to be composed of the given topic.

Our LDA in Chapter 4 identified several topics in our corpus of posts that we would like to "guide" our new model for this study toward discovering. We will use knowledge of our data gathered from Chapter 3 to refine the topics we are looking for into specific groupings that we have identified as important in our documents. Our overall goal in this study is to provide our logistic regression model with sufficient information about each document so that it can give us as accurate an assessment as possible of the association between topics and emotional reactions. But what exactly constitutes "sufficient" information in this case? We address this problem of sufficiency in two ways: first, we increase the specificity of the seed topics we give to the LDA, and then we significantly raise the overall number of topics we tell our LDA to find. In the following paragraphs, we describe the process of choosing and refining our topics.

5.1.3.2 "Sufficient" Priors

To achieve granular, hyper-specific topics, we divide scientific fields into subfields and general document categories into sub-categories that contain highly specialized and related word groupings. As an example, biology is split from a single topic into multiple topics: one for words such as "cell," "molecule," and "protein"; one for words such as "genetics," "DNA," "genome," and "evolution"; and another for "brain," "neuron," and "neurology." Dividing topics in this way allows us to differentiate between documents that talk about different subjects within a single field of science. A user might feel one way about research related to the brain and another about genetics research. We want to capture this effect in our model. We also want to distinguish between topics that often appear together, but that we know are distinct. An example of this is documents related to vaccines. Often in our dataset, we have seen that vaccines are mentioned with references to autism. We want our model to be able to distinguish between when a post is about vaccines only, and when a post is about vaccines causing or not causing autism. Statistically, we might expect these two topics to appear together; an unsupervised LDA is likely to associate these words into a single topic. Our guided approach gives us more specificity in modeling our documents.

We have also included topics for general words, such as general negative words (e.g., curse words, abusive language, "violence," "harassment," "pain," and "bad"), positive words

Table 5.1: Seed topics for our LDA by category.

Topic Group	#	Subtopics	#	Associated Words
Natural Sciences	41	Environmental Science	11	climate, climate change, global warming, pesticides, fracking, mining, fossil fuels, greenhouse gas, ice caps, conservation
		Plant/Animal Science	13	animals, plants, trees, deforestation, extinction, endangered, insects, national park, seeds, veterinary science
		Physics/Cosmology	10	universe, solar system, particle, quantum, star, sun, moon, planet, rocket, orbit, astronaut
		Biology/Chemistry	5	chemical, cell, molecule, protein, brain, neuron, genetic, gene, DNA
		Geology/Geography	2	sediment, igneous, rock, geological, geographic
Formal Sciences	6	Mathematics	1	math, equation, geometry, discrete, algebra
		Statistics	1	statistics, probability, distribution, stats, stat
		Computer Science	4	AI, algorithm, deep learning, neural networks, automation, engineering, robotics, machine
Health Sciences	31	Medicine & Health Care	12	doctor, medical, healthcare, sick, hospital, surgery, illness, cancer, HIV, stem cells, blood, bleeding, cardiology
		Vaccines	3	vaccination, immunization, anti-vaxxer, autism
		Mental Health	5	psychiatry, mental illness, mental disorder, dementia, Alzheimer, suicide, depression, anxiety, loneliness
		Nutritional Science	5	diet, sugar, obesity, vegan, allergies, gluten-free
		Pharmacology/Drugs	6	pharmaceutical, prescription, opioids, marijuana, cannabis, THC, cigarettes, tobacco, addiction, cocaine
Social Sciences	19	Arts & Humanities	8	history, art, culture, literature, painting, ethics, philosophy, logic, linguistics, language, feminism, archaeology
		Psychology	1	psychology, psychologist, psychological
		Economics	5	economy, inequality, budget, fiscal, finance, welfare, poverty, recession, tax, homelessness
		Religion	5	god, bible, Christianity, sin, Buddhism, Islam, Hinduism
Government & Politics	20	Politics (general)	8	politician, policy, national, senate, parliament, citizen, immigration, vote, conservative, liberal, elections
		Politics (specific)	5	Environmental Protection Act (EPA), Trump, Merkel, prime minister, president, Brexit, ACA, guns, United States, America, Europe
		War	2	war, United States, president, prime minister, military, nuclear war
		Law and Law Enforcement	5	law, legal, court, judge, lawyer, illegal, police, prison, crime
Identity	19	Gender	3	male, female, masculinity, transgender, boy, girl
		Marriage	1	marry, marriage, husband, wife
		Sexuality	13	sex, reproduction, homosexual, bisexual, heterosexual, abortion, contraceptives
		Diversity & Discrimination	2	diversity, sexism, racism, racist
General Topics	28	Academia	5	PhD, academic, university, college, graduate, student, school, find, link, factor, role, system, methodology, peer review, journal, study, insight
		Death	2	kill, murder, homicide, execute, die, death, dead, obituary
		Negative Words	12	[explicit language], disaster, pain, hurt, harm, torture, horror, abuse, violence, fool, crazy, stupid, idiot
		Positive Words	2	good, great, epic, neat, safe, safety
		Emotional Categories	7	happy, love, joy, excited, amazed, funny, laugh, angry, sad, cry, fear, afraid, scared, hate
Total seed topics:	164			

(e.g., “good,” “great,” “safe,” and “neat”), and general emotional categories (e.g., “happy,” “sad,” “joy,” “love,” “cry,” “fear,” and “hate”). In total, our LDA has 164 seed topics that we want our model to identify. Table 5.1 shows the broad categories of scientific topics we have used for this study, along with the number of seed topics in each category that we have specified. These topics fall into seven broad categories: the Natural Sciences, Formal Sciences, Health Sciences, Social Sciences, Government and Politics, Identity, and General Topics. The structuring of the subtopics within these broad topic groups is carried out somewhat intuitively based on experience looking through the vocabulary of the corpus, as well as in spending time exploring posts in Chapter 3. The overall structure of these topics and subtopics are shown in Figure 5.3. The leaf nodes on the left branch from the root represent the 164 topics we seeded; the leaves on the right side of the tree represent the 836 randomly discovered topics our LDA model found. In all, we asked our LDA model to discover 1,000 topics, but only provided the prior distribution for 164 of those. The 164 seeded topics we used are given in Appendix B.

The second measure we took to address our “sufficiency” problem was significantly raising the total number of topics we asked our LDA to find. Here, we borrowed one of the core assumptions of LDA—namely that documents represent “infinite mixtures” of underlying topics. We are interested in a handful of specific topics, but these only make up an infinitely small fraction of the topics that are represented in our documents. Intuitively, we can follow this line of reasoning by imagining that we want our seed topics to represent only a small portion of the total topics our model identifies in the corpus. Increasing the number of topics we searched for to 1,000 decreased the likelihood that the model would associate other, potentially unrelated words with our seed topics. Spreading the probability out across more topics increases the purity of the resulting topics, which is beneficial to us because it allows us to assess the emotional reactions associated with specific topics with more certainty that other words or topics are not influencing these reactions.

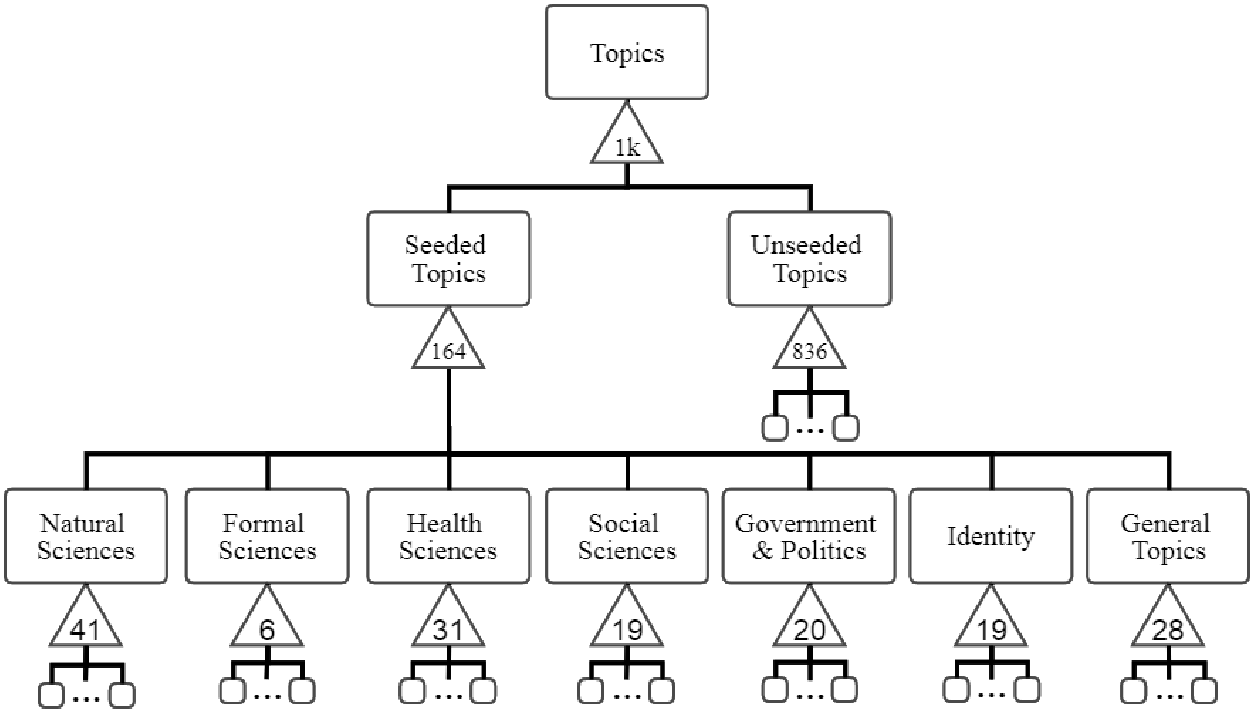


Figure 5.3: A tree diagram of our topic hierarchy. The number of topics in each of the categories is displayed in the triangles beneath the topic nodes.

5.1.3.3 Why LDA?

We have provided a fair amount of detail about how LDA topic models work and how we use them for our study, but we have not yet considered the question of why we would want to use them in the first place. We have mentioned that the use and exploration of topic models in NLP began in the 1990s. The advent of LDA led to an explosion of research in different directions, finding a variety of useful ways to extend the core concepts of LDA [106, 110], to speed up or optimize its implementation [111], and to find new and surprising applications for LDA. This field of research has slowed down in the past few years, primarily due to the rise of a different form of document representation: text embeddings.

Text embedding has captured the attention of researchers and the general public alike, wowing them with generative models such as the bidirectional encoder representations from transformers model (BERT) developed by Google AI [112], which—among other things—can create surprisingly realistic texts. Text embeddings work at different levels (word, sentence, paragraph, document) and convert a set of words into a vector—sometimes of a set number of dimensions (as is the case with BERT), sometimes not—of continuous values, mapping the meaning of a given textual level into Euclidean space. Embeddings have proven useful to researchers for the mathematical properties of these meaning-encoding vectors.

Making inferences about meaning using embeddings is, however, somewhat difficult. One can tell, for example, that two sentences may have a similar meanings based on the Euclidean distance between their vector representations; but one may not know in what ways they are similar, or even what the documents are about from only looking at the vectors. It is primarily for this reason that we have elected to use LDA topic models to represent our texts over embeddings. We prefer the interpretability of LDA—especially given we are using our prior knowledge about our documents to seed topics into our model. We prefer to represent our documents through probabilistic rather than deterministic methods.

5.1.4 Data Analysis

5.1.4.1 Filtering and Formatting Data

As we have done in previous studies, we want to choose data that will best help us accomplish our goal of identifying emotions toward topics of science. One problem with Dataset A is that a minority of posts receive a majority of the reactions. We saw in Table 4.1 that nearly a quarter of all posts did not receive any clicks at all, and that while about 75% of the posts received one or more “Likes,” only a small share of those received any of the special reactions. We also found that the emotional connotation of a “Like” depends in large part on the context in which the reaction is applied, i.e. “Likes” can have a meaning that is positive in some situations and negative in others. For this reason, we decided only to use the five special reactions as the response variables in our regression model, and to exclude “Likes” from consideration. Focusing only on the five special reactions improves our ability to identify emotions toward science as expressed on public Facebook pages.

We filtered Dataset A to only those posts that received more than five total special reactions, and then eliminated posts that were primarily non-English and any non-English words from the posts that remained. We then removed hyperlinks and email addresses from the messages that remained. We further limited our subset to only those posts that contained more than three words (e.g., posts such as “So cool!” or “Neat” were removed). In looking at our subset of data, we found that this was an appropriate threshold at which posts started containing some details about the specific topic of science that was being shared. We were left with 10,931 unique posts that received a total of 436,629 special reactions.

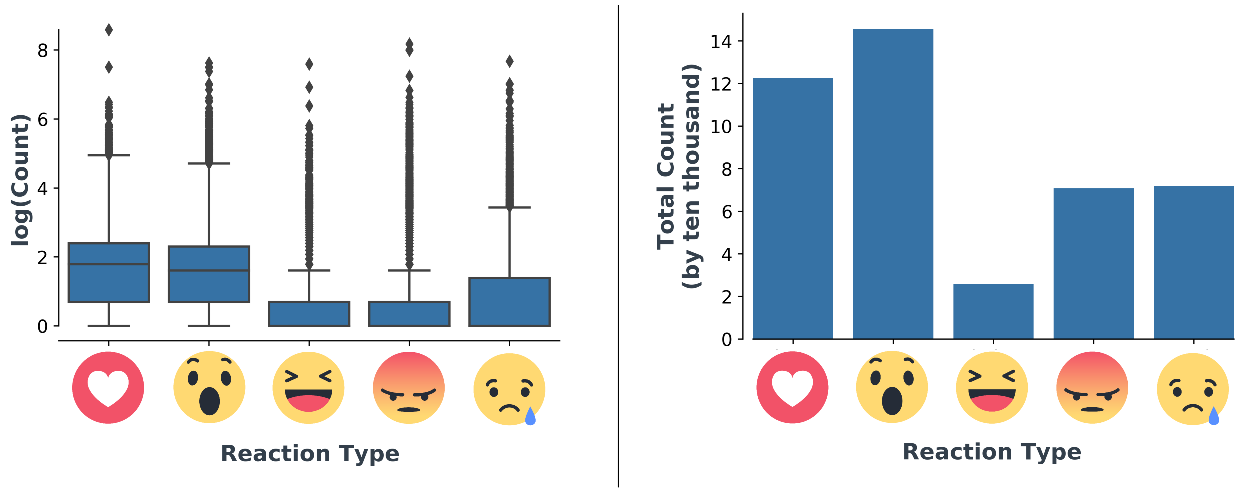
Facebook posts with URLs often generate an image from the hyperlink. This image is often taken from the page the post links to, and is more often than not accompanied by the title of the article that is being shared. This makes it so that someone who views a post will not only see the text of the post, but also will see the title of the article in the link. Studies have shown that click-through rates are low on social media [113], but we can at least say that a person who sees the post will immediately be able to see the title of the article shared. We therefore combined the text of the post with the title of the article to more accurately represent the text visible to respondents. We call this combination of post text and title a “document.”

We wanted the emotional responses to the documents to be represented as categorical reactions not counts. To accomplish this, we transformed our subset of posts into a new format with a set of independent features X that represent the document and a single dependent response feature Y, where $Y \in \{love, wow, laughter, sad, anger\}$. This transformation produced a collection of 436,629 unique responses to our documents. The average number of special reactions received by a document in our transformed dataset was 39.94 with a standard deviation of 137.37.

It is worth emphasizing the fact that in this new form, the set of documents in X are not all unique, and each unique document in X may receive different responses Y—in other words, document d_1 may have received a "Love" and a "Wow" reaction, which would produce two entries in our data, one where $X = d_1, Y = love$ and another for $X = d_1, Y = wow$. At first glance, this idea that the same document can produce different reactions may seem puzzling. How can we expect our model to learn what topics in a document produce a specific emotional response when it is receiving conflicting information? How can one specific combination of features produce different responses? These questions touch on two very important ideas that we have tried to grapple with in this study. First, social-media data is noisy. It is difficult to capture enough context to account for small changes in user responses. Our best hope is to give our models as much information as we possibly can about the phenomena we want to study. Second, we want our model to capture the *average* effect of changes in our documents. In answer to the first idea, we have shown that by increasing the number of seed topics while at the same time reducing the proportion of seed topics to overall topics, we increase the amount of context we are able to provide to our model. On the second point, it is enough to say that measuring average effects is exactly what logistic regression does.

5.1.4.2 Distribution of the Response Classes

Figures 5.4a and 5.4b show the distribution of reaction types in the data we used for this study. We see from the boxplot in Figure 5.4a (which is shown on a logarithmic scale



(a)

(b)

Figure 5.4: A boxplot (a) and bar chart (b) for the reactions in the data.

to improve visibility of the distribution) that the distribution of reactions per post is right-skewed for all categories. “Love” and “Wow” reactions were more common than the other types, and their median values are significantly higher and their interquartile range (IQR) is entirely non-zero. The median number of “Laughter,” “Anger,” and “Sad” reactions are all zero, but these distributions also have long tails that span almost as far as those of the more common “Love” and “Wow” reactions. “Sad” reactions were applied to more posts, but “Anger” and “Laughter” reactions tended to be higher where applied. Figure 5.4b shows the total number of each reaction type applied to all posts in our set. We notice that though the median and IQR for the “Love” reaction are slightly above those of the “Wow” reaction, the overall number of “Wow” reactions is higher. This tells us that though “Love” reactions are applied more frequently and when “Wow” reactions are applied they are present in higher numbers.

Knowing the distribution of our response variable is important because we want to ensure that the classes are relatively balanced. If there are not enough examples of a given class, models will have difficulty predicting responses of that class and will often over-predict the dominant classes. A representative example of this is shown in the problem of spam detection in emails [114] and other forums such as Twitter [115]. In this problem, a classifier must decide which emails (or, more generally, documents) are spam and which are not. Generally, examples of spam occur less frequently. Training a model on a “natural” dataset with very few examples of spam results in a model that will always predict that emails are not spam. Such a model will achieve very high accuracy without actually solving the problem we are interested in. Several remedies exist for this, including over-sampling examples of spam email in a training set, under-sampling examples of non-spam, using synthetic minority oversampling technique (SMOTE) to generate new examples of spam from known cases [116], as well as more novel techniques of synthetic oversampling using the majority class [117].

“Love” reactions represent 28% of the data with 122.4 thousand examples, “Wow” 33.4% with 145.7 thousand, “Laughter” 5.9% with 25.8 thousand, “Anger” 16.2% with 70.8 thousand, and “Sad” 16.5% with 71.8 thousand. The classes are not evenly distributed, but there are still a significant number of examples in each class for a logistic model to discriminate between them. We should bear in mind that training our model with data as is (i.e., not over- or under-sampled) will yield a model that is more likely to assign a larger portion of the total probability to the larger classes “Love” and “Wow.” But learning the “natural” distribution of these reactions is a feature of our experiment not a bug. We want our model to predict the probability that an average user will respond a certain way to a given composition of topics, not for it to more accurately identify when they may give a rarer response. For this reason, we have chosen not to apply over- or under-sampling with our data before training our model, and instead to use the whole of our data for modeling.

5.1.4.3 Choosing a Reference Class

We have mentioned that multinomial logistic regression produces a model that measures the log odds of the probability of responses against a baseline response variable, which is often called the *referent* or *reference* class. The choice of referent does not affect the output of the model in that the probabilities estimated by the model will be the same regardless of the baseline used. We can recover the probabilities estimated by our model for each of our independent features by iterating through the response classes using the following transformation:

$$\begin{aligned}
&P(y_i = 1) = \frac{e^{\hat{\beta}_1 x_i}}{1 + \sum_{k=1}^{K-1} e^{\hat{\beta}_k x_i}} \\
&P(y_i = 2) = \frac{e^{\hat{\beta}_2 x_i}}{1 + \sum_{k=1}^{K-1} e^{\hat{\beta}_k x_i}} \\
&\ldots \\
&P(y_i = K-1) = \frac{e^{\hat{\beta}_{K-1} x_i}}{1 + \sum_{k=1}^{K-1} e^{\hat{\beta}_k x_i}}
\end{aligned}$$

and then finding the final probability associated with our referent class:

$$P(y_i = K) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\hat{\beta}_k x_i}}.$$

We have used the “Anger” class as referent. This choice seemed to produce the most meaningful log-odds ratios of all combinations, directly comparing positively valenced reaction types such as “Love” and “Wow” to a negatively valenced reaction, and also drawing out subtle differences in the uses of “Sad” and “Anger” reaction types. In our results section, we will present both the coefficients produced by our model (i.e., the log odds of seeing each reaction type versus seeing an “Anger” reaction) and these probabilities of each class.

5.1.5 Compositional Data & the Multicollinearity Problem

Using topic distributions to represent the text of posts in our dataset presents a unique challenge. Statistical models are often built on the assumption that the explanatory variables are independent of each other. One of the main reasons for this is that when we try to measure the effect these variables have on a response, having correlations makes it difficult or impossible to measure how a change in one of the explanatory variables influences a change in the response. All of the independent features should be free to vary without affecting the values of the other explanatory features. Correlations obscure the inferences we want to make about relationships in our data, and as such it is common to remove one of a pair of correlated features.

Compositions are data that is made up of related parts or components, all of which form a proportion of a whole. They are used to represent data in many domains, such as the chemical makeup of compounds, the proportions of different minerals in rocks, or the equity allocation in financial portfolios. Compositions require special care for analysis, as the individual parts do not carry absolute information. For example, if we are told that students identifying as women represent 49.7% of the population attending a university, we have received no information about the total number of people attending that university—the best we could possibly say is that with that level of precision, and given the value is not rounded, there are at least 1,000 students enrolled at the university. Knowing that the proportion of women also tells us that people who do not identify that way make up 50.3% of the university's population. This latter inference cannot be made if instead we are told that there are 11,427 students who identify as women. The number 11,427 carries absolute value and is not dependent on any other information about the population.

All this seems relatively intuitive, but identifying that a set of data contains relative rather than absolute values is an important step that should influence the methods we use to analyze our data and make inferences with it. A theoretical system for this special class of data was set in place with the 1982 publication of John Aitchison's *The Statistical Analysis of Compositional Data* [118]. Extending Aitchison's practices and theoretical system has been taken up in the past decade-and-a-half by others, who have developed effective methods for using compositional data with common statistical methods such as regression modeling, principal component analysis, and exploratory data analysis [119, 120, 121]. Appendix A offers a more extensive and technical look at the properties of compositions, and some of the ways that researchers have used them in their analyses.

5.1.5.1 Isometric Logratio Transformation

In short, the problem of compositional data stems from the fact that compositions in D dimensional space occupy only a subset of $D-1$ dimensions known as a *simplex*, which is a generalized notion of a triangle in n-dimensional space. The first theoretical framework for handling of compositions was proposed by Aitcheson [122]. To move compositional vectors out of the simplex into Euclidean space, he developed two transformations: *additive logratio transformation* (alr) and *centered logratio transformation* (clr). But these methods lead to further problems in modeling with compositional data and still produce singularity—a result of clr, for example, is that all features produced by the transformation sum to zero. To solve this problem, Hron et al. [120] produced a new type of weighted transformation they called *isometric logratio transformation* (ilr), which moves a composition out of a simplical subset in $\mathbb{R}^D$ into a new space of $\mathbb{R}^{D-1}$ on the orthonormal basis of a selected feature x_i. In using x_i as the basis for ilr, this feature becomes the "vanishing dimension" around which all other features—now represented by a feature vector $\mathbf{z}$—are configured in the new space. We can then use this new vector $\mathbf{z}$ in a relatively straight-forward manner to make inferences about the effects of x_i on the response.

Performing ilr involves sequential binary partitioning of the data into a left vector $\mathbf{l}$ and a right vector $\mathbf{r}$. Equation 5.4 demonstrates the process by which a vector $\mathbf{x}$ can be transformed into a new vector $\mathbf{z}$ through this process of iterative partitioning.

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \log(\frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}), i = 1, ..., D-1 \tag{5.4}$$

Each feature in $\mathbf{x}$ is "folded into" the geometric mean of $\mathbf{r}$ and the logarithm is taken of this quotient. Each logratio is then weighted by a factor of $\sqrt{\frac{D-i}{D-i+1}}$, which increases the spread of the coordinates in the new space. We can then measure the effect that changes in x_i have on the response variable by using $\mathbf{z}$ as input into a model and regarding the coefficients computed for the first feature z_1. Performing ilr on the basis of x_i for each $i = 1, 2, ..., D$ and using the resulting features $\mathbf{z}$ as input to a model gives us the set of coefficients $\hat{\beta}$ that we expect from a logistic regression model, and we can interpret these in a relatively straight-forward manner [120]. Because ilr transformations are orthogonal transformations of each other [119, 120], the fit of the resulting models (i.a., their log-likelihood, their values of "pseudo r-squared") will be invariant, with the difference being the values of $\hat{\beta}$.

Figure 5.5 shows what this process looks like when applied to a feature vector $\mathbf{x}$ in a tree-based representation, where ilr is performed using a composition (x_1, x_2, x_3, x_4) and producing a transformed vector (z_1, z_2, z_3) on the orthonormal basis of the first variable (x_1). The 4-dimensional space is reduced in the process to three dimensions. Sequential binary partitioning divides the components at each level into a left and right subtree, with the left containing the current component and the right containing the remaining parts of $\mathbf{x}$.

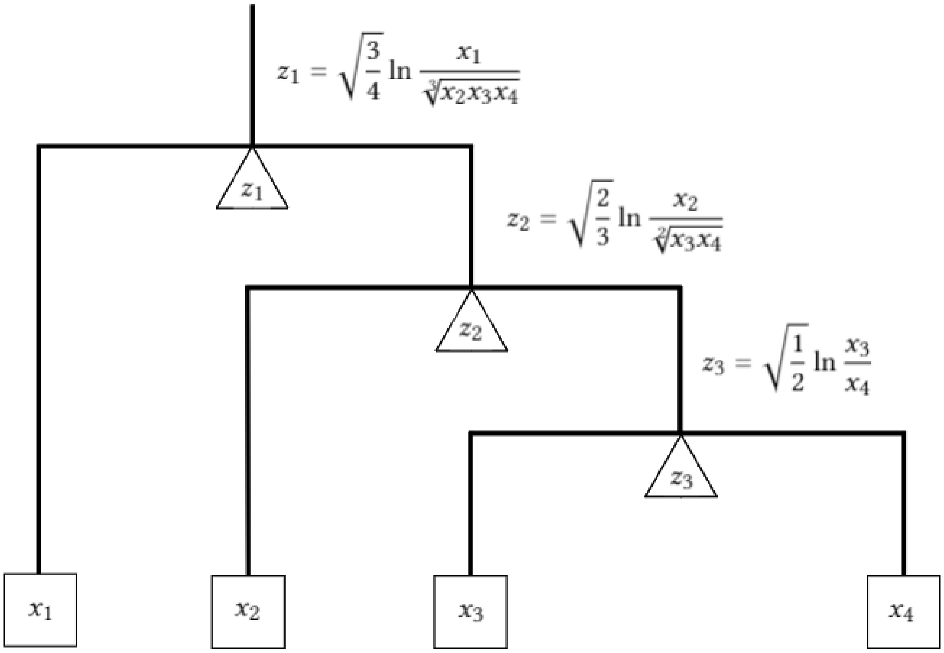

$z_1 = \sqrt{\frac{3}{4}} \ln \frac{x_1}{\sqrt[3]{x_2 x_3 x_4}}$
z_1
$z_2 = \sqrt{\frac{2}{3}} \ln \frac{x_2}{\sqrt[2]{x_3 x_4}}$
z_2
$z_3 = \sqrt{\frac{1}{2}} \ln \frac{x_3}{x_4}$
z_3
x_1
x_2
x_3
x_4

Figure 5.5: A balanced tree demonstrating ilr transformation.

5.1.5.2 Interpreting Regression Coefficients

It is worth expanding a bit on how we interpret the regression coefficients ($\hat{\beta}$) produced by our model. As shown earlier in our introduction to logistic models, log odds can be transformed into odds that represent $\frac{p(X)}{1-p(X)}$. This transformation is monotonic, which is to say that the greater the log odds, the greater the odds and vice versa. An odds ratio can then be monotonically transformed into a probability. Log odds are continuous values ranging from $-\infty$ to $+\infty$. Positive values can be interpreted as showing that the probability of a success is greater than the probability of failure. The larger the log odds are, the larger the probability of success. Conversely, the more extreme the log odds are toward negative infinity, the higher the probability of failure. A log odds of zero indicate that the probability of success is equal to the probability of failure.

In our logistic model, we have selected the "Anger" response as the reference class. For all topics, our model will give us four values of $\hat{\beta}^k$—one for each of the remaining special reactions. Each of these values represents the log odds of seeing that reaction k over an "Anger" reaction. For example, if our model estimates that the coefficient for topic i and reaction k is $\hat{\beta}_i^k = 0.64$, it has predicted that the chance of seeing reaction k chosen over an "Anger" reaction increases with the increased presence of the topic i. The specific amount we expect to see the log odds change with topic i requires a bit more explanation. The coefficients indicate the amount we expect the log odds to change when a given independent variable increases by one unit. Understanding what a unit change is usually relatively simple, but because we have used isometric logratios to transform our features, our unit measurement is no longer as straight-forward as "1%" or "1 unit of probability": it is now "one unit change in the isometric logratio value."

To grasp what this means, we need to look at how ilr transformation is performed. In the following discussion, we are careful to distinguish between a logratio, which is a general term that refers to the logarithm of a ratio, and the isometric logratio, which is a specific type of weighted logratio that we denote by *ilr*. The equation for ilr transformation is given in the following form:

$$z_i = \sqrt{\frac{D-1}{D}} \log\left(\frac{x_i}{\sqrt[D-1]{\prod_{j \neq i} x_j}}\right) \tag{5.5}$$

where D is the number of dimensions, which in our case is the number of topics. Equation 5.5 shows how the first feature of vector $\mathbf{x} \in \mathbb{R}^D$ is transformed into the first feature of a new vector $\mathbf{z} \in \mathbb{R}^{D-1}$. The value of x_i is divided by the geometric mean of the remaining parts of $\mathbf{x}$, and then the natural logarithm of this quotient is taken. This value is then weighted by the quantity $\sqrt{\frac{D-1}{D}}$. Our logistic model's $\hat{\beta}_i^k$ shows the expected change in the log odds of response k when z_i increases by one unit. So what does it mean for z_i to grow by one unit?

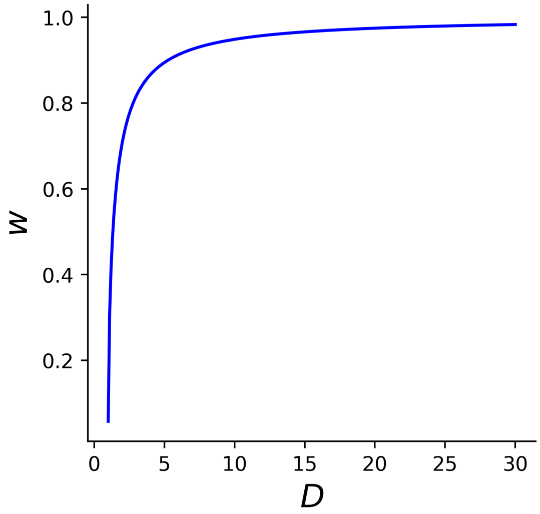


Figure 5.6: A plot of the weighting coefficient w as a function of D.

Let us start by examining the weighting factor, which we define as $w_D = \sqrt{\frac{D-1}{D}}$. The behavior of this quantity is shown in Figure 5.6 where the value of w is plotted as a function of D. We see that as D approaches infinity, w approaches 1. By $D = 30$ this value is already very close to one. By $D = 164$ (the number of features in our dataset) w is approximately 0.997. Therefore as D grows, w reduces our overall quantity by less and less. We can give the following definition about the behavior of the quantity w:

$$1 = \lim_{D \to \infty} w_D \tag{5.6}$$

The quantity expressed by $\log(\frac{x_i}{\sqrt[D-1]{\prod_{j \neq i} x_j}})$ is more interesting because it is dependent on the specific values of $\mathbf{x}$. Both linear and logistic regression measure the effect that changes in one independent feature have on the response variable while all other features are treated as constants. The quantity in the denominator of this equation represents the geometric mean of all features $x_{j \neq i}$, i.e. all the average of all features we are holding constant while

x_i varies. Following this idea, we can reduce the complexity of our problem by treating this mean as a constant, which we denote as c. We want to quantify the change in x_i that will produce a one-unit change in a logratio, so we give this unknown quantity a name k and express it in the following terms:

$$\log(\frac{x_i + k}{c}) - \log(\frac{x_i}{c}) = 1 \tag{5.7}$$

Simplifying this equation, we see that k depends on the value of x_i:

$$k = x_i(e - 1) \tag{5.8}$$

where e is Euler's number. Now that we have an expression for k in terms of x_i we can substitute it into the term expressed by the left logratio's numerator in Equation 5.7 to see that:

$$x_i + x_i(e - 1) = ex_i. \tag{5.9}$$

This statement tells us that when x_i increases by a factor of e the logratio increases by 1 unit:

$$\log(ex_i) - \log(x_i) = 1. \tag{5.10}$$

Additive increases in the logratio correspond to multiplicative increases in x_i. This is a good finding (though not entirely surprising given that we are working with logarithms). But we still need to combine this information with what we concluded about our term w. We want to define the overall behavior of an ilr transformed feature. We take the same approach as we did above and give the unknown quantity of interest the name r, which we define as the factor by which x_i must increase to increase z_i by one unit. We know that $r = e$ in the case when D is infinitely large and $w = 1$, but what happens when this is not the case? We can rewrite Equation 5.5 with the constant terms removed (the geometric means cancel each other out) and our weights added back in as follows:

$$w \log(rx_i) - w \log(x_i) = 1 \tag{5.11}$$

We can put our conclusion about w together with what we discovered about the logratio to arrive at a formal definition for what is needed to increase z_i by 1 unit: when D goes to infinity, w goes to 1, and when w goes to 1, r approaches e from the right. When the value of D decreases, the value r grows from e toward infinity. We can formally define this relationship as follows:

$$r^w = e. \tag{5.12}$$

Equation 5.12 gives us the power to interpret our regression coefficients. We conclude that $\hat{\beta}_i$ represents the amount the log odds are expected to change when x_i is increased by a factor r, a value that approaches e from above as D goes to infinity. In our study, we can say that $r \approx e$ (it is actually a value 0.3% greater than e). Unfortunately this is as good an interpretation as we can get, and this is the price we pay for using compositional data to represent our documents.

5.1.5.3 Why ilr?

Finally, we might ask also if it is necessary to perform ilr at all. A naïve alternative follows from the idea that because we have a singular vector $\mathbf{x}$ where all n parts sum to 1, can we not just remove any part and use the other $n-1$ features without any transformation? As Hron et al. demonstrate [120] with a proof through application, acting on this idea results in a model with unreliable inference statistics that will vary depending on which feature x_i is excluded. We might still be able to produce a model with equivalent predictive powers, but it would be a black-box—torpedoing the primary reasons we have used regression models in the first place.

5.1.6 Putting It All Together

This study puts together many of the ideas that have been developed earlier to accomplish one of the primary goals of this thesis: can we use our Facebook data to learn how users feel toward topics in science? Figure 5.7 shows the main outlines of the steps we take in this final study. It shows how we change our representation of the texts using our LDA, how we re-figure our data so that we have categorical responses, and how we use multinomial logistic regression to estimate the effect that the presence of certain topics has on the probability of seeing certain reactions. The following sections address several basic questions that may remain about why we have made the decisions we have in this study. Answering these questions will help us better understand what we are doing in this study, and how the choices we have made at each step in the process help us accomplish our goals better than alternative approaches we might have taken. We also include specific details about how we performed our study—the languages and tools we used, parameters of models, and the specific steps taken at important junctures.

5.1.6.1 Implementation Details

As in Chapter 4, we use Gensim's LDA model implemented in Python in this study. To address the problems with the compositional nature of our data, we have built a wrapper class, which we call a *CoDa* (short for "Compositional Data"), around a Pandas [123]

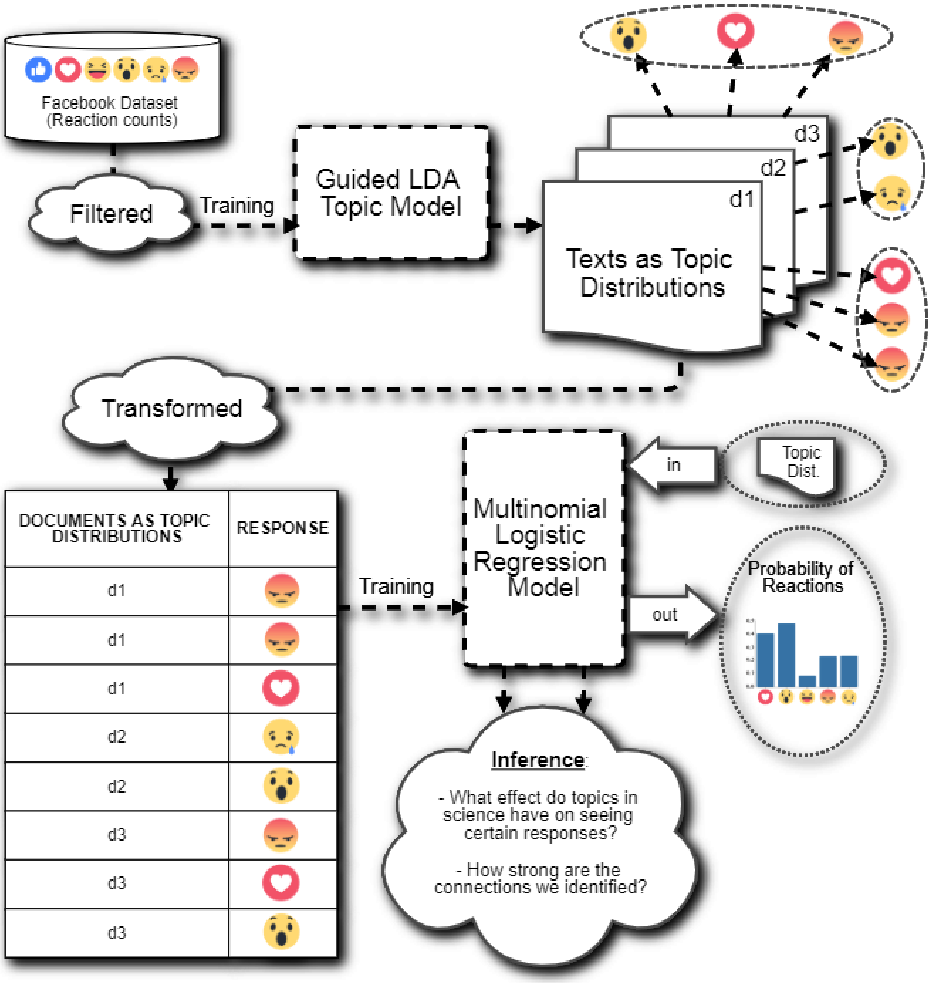


Figure 5.7: Workflow diagram for this study.

DataFrame data structure, with a fast implementation of isometric logratio transformation. We built another wrapper class we call the *CMNLogit* ("Compositional Multinomial Logistic Regression") over an implementation of a multinomial logistic regression model from the Python statsmodels library [124] that takes a CoDa object as an exogenous (independent) variable. This class uses the CoDa object and performs the necessary transformations, and stores the results in a convenient format similar to the one provided with the models from statsmodels.

As the optimization function for our multinomial logistic regression model, we use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, which is an iterative method for solving unconstrained nonconvex optimization problems [125, 126, 127]. We add an intercept to our feature vectors, a term that represents the marginal distribution of the response variable. To compute the coefficients for each of our independent variables, we use ilr transformation on the orthonormal basis of a given feature and then train a logistic model with those transformed features. We take the resulting coefficient, standard error, z-score, and p-value for the first variable, which represent the effect of the independent variable that is the basis of the transformation. We perform this process iteratively, taking each feature as the basis of transformation and then storing the result. As we have mentioned, the overall model performance (i.e., the log-likelihood of the model, the psuedo-R-squared value testing "goodness-of-fit," the predictive accuracy of the model—assuming new observations are transformed along the same basis as the data used for training) will not change based on which feature is used as the basis for ilr transformation; the estimated coefficients will, however, be different from one basis to another. In other words, to compute all our coefficients we need to fit D logistic models, each one using a given feature n as the basis for transformation.

We only used a *subcomposition*, or a selected subset of a composition, which is then treated as a composition, of the 164 seeded topics to train our multinomial models. The 836 unseeded topics are "sparse" random variables, i.e. they represent thinly spread distributions over a great number of words of which we are nominally not interested in measuring the effects. These unseeded topics are highly covariant in that they are all near zero for almost all documents. MLE determines that these features are linearly dependent and cannot converge on a solution because of the singular nature of the data. As shown above, one of the assumptions of a logit model is that there is little or no multicollinearity among the features. Including the unseeded topics in our feature vector $\mathbf{x}$ would violate this assumption. Because of the large number of topics and the problem with floating-point numbers, we normalized our topic composition to a basis of $k = 1,000$. This lowered the possibility of rounded zeros in our data, which can be problematic for MLE. Altering the value of k does not change the value of the coefficients.

5.2 Results

5.2.1 Analysis of Topics

We were interested to see how the proportion that each topic represented varied across all documents. Table 5.2 shows twelve topics (a topic here is a leaf node in the diagram presented in Figure 5.3) that received among the highest mean values and twelve that received the lowest mean values in our data. We have used the geometric mean and standard deviation instead of the arithmetic mean and standard deviation because of its robustness in dealing with proportional data. The value of a given topic represents the proportion of a document that we estimate may be generated by that topic. The values for μ and σ in the table are given as percentages (out of 100) instead of proportions (out of 1.0), e.g. topic 101 having $\mu = 3.218$ means that it occupied 3.218% of a document on average. The table also shows the maximum and minimum values for each topic presented.

These statistics are important to us for several reasons. First, it is useful to know which topics are most represented in our documents. A high value of a particular topic means that it produces a large proportion of the words in that document. We see in the table that the most represented topics are from our *General Topics→Academia* categories. Words such as "peer review," "journal," "study," "find," and "factor" appear most frequently in our set. This is not a surprising outcome. Our posts are all related to research findings and the posts themselves often present scientific conclusions as the result of a new study or journal article.

Table 5.2: A selection of the highest and lowest values of topic geometric means μ along with their geometric standard deviations σ, as well as the maximum and minimum values. Values have been represented as units per hundred to manage small decimal values.

Topic		**Statistics**			
Topic No.	Keywords	μ	σ	max	min
	High Mean				
101	peer review, journal, paper, study, methodology, evaluate	3.218	1.820	38.385	0.233
98	find, could, link association, factor, role	3.174	1.859	27.906	0.273
47	health, healthy, medicine, medical	1.612	2.099	34.935	0.109
102	politics, government, nation, politician	1.495	2.117	34.771	0.035
51	sick, sickness, illness, disease, infection, epidemic, virus, bacteria	1.211	2.162	36.805	0.060
97	PhD, doctoral, academic, academia, professor, university, school, student	1.180	2.171	37.009	0.043
99	system, structure, organize, method, process, approach	1.137	1.947	27.657	0.069
36	cell, molecule, protein, molecular, enzyme	1.068	2.166	38.351	0.049
155	good, great, neat	0.919	1.886	20.590	0.024
0	environment, environmental, nature, natural	0.904	2.008	31.417	0.033
50	hospital, patient	0.888	2.075	44.485	0.020
37	genetic, DNA, genome, gene, mutation, evolution	0.711	2.152	31.824	0.027
	Low Mean				
25	cosmology, cosmological	0.007	1.698	12.513	0.000
132	transgender, trans	0.006	1.712	14.301	0.000
83	unethical	0.006	1.745	4.768	0.000
135	contraceptives, condoms	0.006	1.682	9.107	0.000
44	automate, automation	0.006	1.684	8.022	0.000
21	euthanasia, euthanize	0.006	1.688	18.204	0.000
91	Hinduism, Hindu	0.006	1.669	25.025	0.000
89	Buddhism, Buddhist	0.006	1.665	13.646	0.000
133	heterosexual, heterosexuality	0.006	1.671	7.698	0.000
131	bisexual, bisexuality	0.006	1.663	19.241	0.000
8	fracking, frack	0.006	1.824	14.301	0.000
137	pornography	0.006	1.656	14.301	0.000

These keywords can appear in articles of all other subsets. For example, a post may present details from a study about the environment, or about history, or about a particular species of ant. All of these examples might use the words “study,” “paper,” or “link.”

The topics with the highest mean values in our set (101 and 98) also have relatively low values of σ. This indicates that the presence of these topics is relatively ubiquitous, and that they generally represent a large portion of the words in a given document. In the high mean category (i.e., the top section of Table 5.2) we also see that topic 155 has a relatively low σ value. Words such as “good,” “great,” and “neat” appear to have lower variation across documents. This fact is not surprising, as these words are general and can be applied across a number of disciplines to describe research.

The next most important categories in the high mean section of the table are related to *Health Sciences*. We saw in Chapter 3 that there were a significant number of pages devoted to health care and medical information. We knew from early analysis of our data that the largest share of articles in our original Altmetrics dataset were related to the health sciences. Seeing this fact represented differently—i.e., as the average value of a set of topics—reassures us that our metrics are capturing phenomena that we would expect to see.

Posts related to *Government & Politics* also represent a relatively large proportion of words in our documents. We saw in Chapter 3 that a significant number of the public pages of science on Facebook were aimed toward using science to bolster politics claims. We also notice the presence of several topics related to biology in the high mean category. Topics related to "cells," "molecules," "protein," "enzymes," "genetics," and "DNA" all fell within the highest twelve means.

Topics that received low mean values represented a broad range of categories: topics about identity ("transgender," "heterosexual," "bisexual"), from the environmental and health sciences ("contraceptives," "euthanasia," "fracking"), and topics relating to several religions ("Hinduism," "Buddhism") all were in this category of low-μ topics. We note that this latter category of religions is possibly the result of our choice to filter out any non-English documents from our set before modeling. It is a good reminder that we should cognizant of the biases we may have introduced into our study in the early stages of data filtering and cleaning, but we note that alternative ways we could have proceeded, such as translating all other languages into English, would have introduced their own biases into our study. It seems that the best we can do is to not overstate our findings and to be as aware as possible of the ways our choices may have influenced the results we see. We are also generally surprised to see topic 44 ("automate," "automation") included in this low-μ category—we might have expected such an ostensibly important topic in this era to represent a more substantial amount of coverage in social-media shares of science.

5.2.2 Model Results

5.2.2.1 Model Statistics

Table 5.3 shows the results of our logistic regression model after training on our documents. The “Anger” reaction was the referent class in our group, so all results are expressed as the odds of seeing a different reaction type (i.e., “Love,” “Sad,” “Wow,” or “Laughter”) versus an “Anger” reaction. Several bits of information are given in the table. For each response category, the top and bottom six topics ranked by the coefficients are displayed.

The standard error represents the level of statistical accuracy of our estimated coefficient. The z-score shows the test statistic, which represents the number of standard errors away from 0 our coefficient falls, and the p-value ($P > |z|$) shows the probability of seeing a value equal to or greater than our test statistic given the null hypothesis that there is no relationship between our independent and dependent variables is true. The final two columns show the 95% confidence interval for our estimated coefficient, i.e. the range in which we are 95% confident that the true value of our coefficients lie. Our p-values are all very close to zero for two reasons: first, we have selected to show only those features whose p-value is less than 0.01, or only those lower than significance level of $\alpha = 0.01$, and second, because the respectable size of our dataset ensures that we are generally highly confident in our results.

There are many interesting details to point out in our table, and it is worth bringing a few up at length. Take for example the first few items in the **Love : Anger** reaction section. Topics 36, 37, and 38 are all under the category *Natural Sciences*, and topic 55 (which is about stem cell research) is closely related even though we have grouped it with *Health Sciences*. These all have a similar high log odds, with research about “cells,” “molecules,”

Table 5.3: Statistics for our logistic regression models.

Topic		**Statistics**				**C.I.**	
Topic No.	Keywords	Coef.	Std. Err.	z	$P > \lvert z \rvert$	[0.025	0.975]
	Love : Anger						
36	cell, molecule, protein, enzyme	0.428	0.015	29.416	0.000	0.400	0.457
37	genetic, DNA, gene, mutation	0.326	0.012	26.207	0.000	0.302	0.351
55	stem cells	0.313	0.028	11.297	0.000	0.259	0.368
38	brain, neurology, synapse, dopamine	0.313	0.014	22.680	0.000	0.286	0.340
14	cat, bird, dog, mouse	0.303	0.009	33.011	0.000	0.285	0.321
101	peer review, journal, study	0.280	0.011	24.818	0.000	0.258	0.302
106	citizen, citizenship	-0.351	0.013	-27.069	0.000	-0.377	-0.326
9	mine, mining	-0.460	0.012	-39.043	0.000	-0.483	-0.437
102	politics, government, national	-0.522	0.008	-63.491	0.000	-0.538	-0.506
8	frack, fracking	-0.525	0.024	-21.450	0.000	-0.573	-0.477
10	pesticide, GMO, Monsanto	-0.526	0.009	-61.646	0.000	-0.543	-0.510
111	Trump, Merkel, president, prime minister	-0.574	0.008	-70.365	0.000	-0.590	-0.558
	Sad : Anger						
160	anger, angry, mad, rage	0.287	0.014	19.830	0.000	0.258	0.315
6	ice, melt, polar, ice cap	0.231	0.009	25.626	0.000	0.214	0.249
66	suicide, anxiety, depression, loneliness	0.227	0.011	20.371	0.000	0.205	0.249
50	hospital, patient, hospitalization	0.215	0.013	16.162	0.000	0.189	0.242
132	transgender, trans	0.213	0.039	5.491	0.000	0.137	0.289
38	brain, neurology, synapse	0.205	0.015	13.319	0.000	0.175	0.235
117	law, legal, court, copyright, attorney	-0.176	0.007	-25.989	0.000	-0.189	-0.163
88	Bible, Christian, preacher, sin	-0.186	0.021	-8.883	0.000	-0.228	-0.145
9	mine, mining	-0.193	0.007	-27.522	0.000	-0.207	-0.179
85	feminism, feminist	-0.193	0.033	-5.794	0.000	-0.258	-0.128
111	Trump, Merkel, president, prime minister	-0.219	0.006	-37.872	0.000	-0.231	-0.208
102	politics, government, national	-0.282	0.008	-37.248	0.000	-0.297	-0.267
	Wow : Anger						
36	cell, molecule, protein, enzyme	0.427	0.014	29.585	0.000	0.399	0.455
38	brain, neurology, synapse, dopamine	0.408	0.014	29.975	0.000	0.381	0.434
37	genetic, DNA, gene, mutation	0.378	0.012	30.768	0.000	0.354	0.402
58	cardiology, cardiac, heart	0.304	0.010	29.848	0.000	0.284	0.324
51	sick, sickness, illness, infection, epidemic	0.290	0.010	27.668	0.000	0.269	0.310
17	dinosaur	0.260	0.027	9.687	0.000	0.207	0.312
94	budget, spending, finance	-0.347	0.008	-44.160	0.000	-0.362	-0.331
12	deforestation	-0.349	0.015	-23.840	0.000	-0.377	-0.320
9	mine, mining	-0.381	0.008	-45.437	0.000	-0.398	-0.365
10	pesticide, GMO, Monsanto	-0.387	0.006	-63.461	0.000	-0.399	-0.375
111	Trump, Merkel, president, prime minister	-0.477	0.007	-66.896	0.000	-0.491	-0.464
102	politics, government, national	-0.643	0.008	-76.366	0.000	-0.659	-0.626
	Laughter : Anger						
71	gluten, gluten free	0.506	0.022	23.354	0.000	0.464	0.549
69	vegan, vegetarian	0.433	0.016	26.576	0.000	0.401	0.465
37	genetic, DNA, gene	0.411	0.015	27.504	0.000	0.381	0.440
144	feces, toilet, bathroom	0.331	0.020	16.234	0.000	0.291	0.371
101	peer review, journal, paper	0.302	0.016	19.333	0.000	0.271	0.333
88	Bible, Christian, preacher, sin	0.280	0.013	20.857	0.000	0.254	0.307
48	healthcare, insurance	-0.313	0.017	-18.409	0.000	-0.346	-0.279
87	religion, religious, god	-0.326	0.018	-18.211	0.000	-0.361	-0.291
10	pesticide, GMO, Monsanto	-0.334	0.014	-24.122	0.000	-0.361	-0.307
47	health, healthy, medicine, medical	-0.353	0.015	-23.334	0.000	-0.383	-0.323
110	EPA, United States, America, DOE	-0.426	0.013	-34.009	0.000	-0.450	-0.401
9	mine, mining	-0.464	0.023	-20.002	0.000	-0.509	-0.418

"proteins," and "enzymes" receiving the highest coefficient. Indeed, topic 36 received the highest log odds of any topic (0.428).

On the other hand, topics in the broad category of *Government & Politics* are generally more likely to inspire responders to use the "Anger" reaction over the "Love" reaction. "Fracking" and "mining," "pesticides" and "genetically modified organisms (GMOs)," and posts about world leaders and politics all have a very high chance of receiving "Anger" reactions over other responses. The strongest log odds in this category (**Love : Anger**) is topic 111 (-0.574). This means when the presents of topic 111 increases by a factor of approximately e, the log odds of seeing a "Love" reaction chosen over an "Anger" reaction will decrease by 0.574. This is a relatively strong association, which is represented by a higher absolute value of the log odds.

The category **Sad : Anger** shows some interesting patterns. In general, we notice that the total range of the log odds is smaller in this category than in any of the others. The ranges of each category are as follows: 1.002 for **Love : Anger**, 0.569 for **Sad : Anger**, 1.07 for **Wow : Anger**, and 0.97 for **Laughter : Anger**. We attribute this significantly smaller range to the correlation in use between the "Sad" and "Anger" reactions we discovered in Chapter 4. There is less overall distinction between these two reactions, which we can see by a smaller spread of log odds across all topics in this category. We can, however, still see that in some topics there is a clear difference in use between these two negatively valenced reactions. Posts about suicide or depression, transgenderism, or posts broadly about the melting ice caps tend to make users more sad than angry. Interestingly, a topic that includes the words "anger," "angry," and "mad" is also more likely to receive "Sad" responses than "Anger." On the other hand, topics about politics, feminism, and Christianity appear to be more likely to provoke "Anger" reactions than "Sad" reactions.

In Chapter 4 we also found a correlation between "Love" and "Wow" reactions. We see that correlation reinforced here with the appearance of very similar topics in the high and low subcategories of both **Wow : Anger** and **Love : Anger**. Topics of biology and genetics have high log odds in both categories, and topics related politics, government, and environmental destruction all have low log odds.

The **Laughter : Anger** category is especially interesting, as until now we have been able to say little of this reaction type relative to the more common types. Here we see that topics related to veganism and vegetarianism, gluten-free dietary science, and a category we will call "bathroom humor" are all more likely to receive "Laughter" responses than "Anger." We notice that topic 71 ("gluten," "gluten free") has the highest log odds of any topic across all categories (0.506). This value indicates that when the value of topic 71 increases by a factor of e, and all other values are held as constant, we expect to see the log odds of seeing "Laughter" reactions over "Anger" will rise by 0.506. Low log odds in this category are assigned to topics regarding environmental destruction, religion, healthcare and insurance, and politics.

We notice that if we were to rank the absolute values of these log odds, those in the negative category would be greater. We interpret this as indicating that though positively valenced reactions such as "Love" and "Wow" are used more frequently than negatively valenced reactions such as "Anger," the relationship between specific topics and this latter category of reactions is stronger—in other words, when negatively valenced responses are given, they are given by a greater share of users than positively valenced responses. Facebook users on these public pages of science appear to have a more visceral negative reaction to political issues, posts related to fracking and mining, and pesticides/GMOs than they have a positive reaction to issues of regarding the natural sciences, or other topics toward which they have positive emotions.

We see this trend in other categories as well. For example, in the category **Wow : Anger**, Topic 102 ("politics," "government") has the lowest log odds (-0.643) of any topic in our data. This coefficient is the most extreme value recorded in our model, indicating that this relationship is very well established. The standard error for that coefficient is 0.008, indicating a high amount of confidence that the true value for this relationship is indeed higher than most other values in our set.

Our broad topic category *Government & Politics* appears most frequently in the low log odds sections. Environmental destruction (mining, fracking, pesticides) and GMOs also appear frequently in the low log odds sections. This indicates that users are more likely to use “Anger” reactions than any other reaction type for posts relating to these topics.

5.2.2.2 Probabilities

Another way we can view our results is as probability distributions for each of the five reaction types. As we showed in the section “Putting It All Together,” we can transform log odds into a probability, and then use the law of total probability to find the distribution of all reaction types per topic. Table 5.4 shows the results of this transformation. For each reaction type, it gives the highest and lowest six topics ranked by probability. For example, the fourth section in the table shows the six highest topics and six lowest topics ranked by the probability of seeing an “Anger” reaction. *Government & Politics* represents the highest of this category, having a 27.2% chance of being associated with an “Anger” reaction, and a relatively low chance of being associated with a “Love” or a “Wow” reaction. Other topics in this high probability section of “Anger” reactions have to do with environmental destruction. We can conclude that our model has determined that posts highly related to political topics and world leaders, as well as those related to topics about the destruction of the environment (fracking, mining, deforestation, pesticides) and genetically modified organisms are the topics with the highest chance of being associated with “Anger” reactions. On the other hand, posts related to animals, biology, hospitals, and genetics are the least likely to be associated with “Anger” reactions. These different topics all have different distributions of the other reactions we can analyze.

Table 5.4: Probabilities of seeing each reaction type for selected topics by high/low values per reaction.

	Topic		Probabilities				
	Topic No.	Keywords	Love	Wow	Laughter	Anger	Sad
		Love					
HIGH	36	cell, molecule, protein, enzyme	0.250	0.250	0.196	0.163	0.140
	55	stem cells	0.249	0.229	0.187	0.182	0.154
	158	amaze, amazement, amazing, amazed	0.241	0.206	0.182	0.195	0.176
	26	black hole, big bang, singularity	0.236	0.234	0.189	0.182	0.159
	41	math, mathematics, geometry, mathematician, algebra	0.234	0.196	0.182	0.204	0.184
	72	marijuana, pot, cannabis, THC, CBD	0.233	0.196	0.237	0.180	0.154
LOW	154	bad	0.161	0.206	0.216	0.215	0.203
	112	Brexit	0.160	0.196	0.210	0.206	0.228
	106	citizen, citizenship	0.152	0.174	0.218	0.216	0.240
	10	pesticide, GMO, Monsanto	0.150	0.173	0.182	0.255	0.240
	8	frack, fracking	0.146	0.175	0.197	0.246	0.236
	111	Trump, Merkel, president, prime minister	0.144	0.159	0.236	0.256	0.205
		Wow					
HIGH	36	cell, molecule, protein, enzyme	0.250	0.250	0.196	0.163	0.140
	38	brain, neurology, synapse, dopamine	0.217	0.238	0.192	0.158	0.195
	17	dinosaur	0.225	0.237	0.198	0.183	0.156
	51	sick, sickness, illness, infection, epidemic	0.203	0.237	0.187	0.177	0.197
	58	cardiology, cardiac, heart	0.220	0.235	0.196	0.174	0.174
	26	black hole, big bang, singularity	0.236	0.234	0.189	0.182	0.159
LOW	95	inequality, welfare, poverty, homelessness	0.182	0.170	0.173	0.231	0.244
	97	PhD, doctoral, academia, professor	0.197	0.168	0.209	0.199	0.226
	94	budget, spending, fiscal, finance	0.167	0.166	0.196	0.235	0.236
	111	Trump, Merkel, president, prime minister	0.144	0.159	0.236	0.256	0.205
	69	vegan, vegetarian	0.185	0.146	0.302	0.196	0.172
	102	politics, government, national	0.161	0.143	0.218	0.272	0.205
		Laughter					
HIGH	71	gluten, gluten free	0.169	0.185	0.303	0.183	0.160
	69	vegan, vegetarian	0.185	0.146	0.302	0.196	0.172
	88	Bible, Christian, preacher, sin	0.190	0.172	0.268	0.202	0.168
	144	feces, toilet, bathroom	0.170	0.188	0.265	0.190	0.187
	85	feminism, feminist	0.206	0.182	0.251	0.198	0.164
	37	genetic, DNA, gene, mutation	0.221	0.233	0.241	0.160	0.146
LOW	9	mine, mining	0.168	0.181	0.167	0.265	0.219
	47	health, healthy, medicine, medical	0.202	0.178	0.164	0.234	0.221
	87	religion, religious, god	0.196	0.177	0.163	0.226	0.239
	23	extinct, extinction, endangered	0.193	0.180	0.161	0.212	0.254
	108	corrupt, corruption	0.214	0.219	0.158	0.195	0.214
	110	EPA, United States, America, DOE	0.193	0.197	0.156	0.239	0.214
		Anger					
HIGH	102	politics, government, national	0.161	0.143	0.218	0.272	0.205
	9	mine, mining	0.168	0.181	0.167	0.265	0.219
	111	Trump, Merkel, president, prime minister	0.144	0.159	0.236	0.256	0.205
	10	pesticide, GMO, Monsanto	0.150	0.173	0.182	0.255	0.240
	8	frack, fracking	0.146	0.175	0.197	0.246	0.236
	12	deforestation	0.171	0.171	0.201	0.242	0.215
LOW	50	hospital, patient, hospitalize	0.204	0.222	0.189	0.172	0.213
	101	peer review, journal, study	0.227	0.205	0.232	0.171	0.165
	14	cat, bird, dog, wolf, horse, monkey, mouse	0.223	0.209	0.211	0.165	0.191
	36	cell, molecule, protein, enzyme	0.250	0.250	0.196	0.163	0.140
	37	genetic, DNA, gene, mutation	0.221	0.233	0.241	0.160	0.146
	38	brain, neurology, synapse, dopamine	0.217	0.238	0.192	0.158	0.195
		Sad					
HIGH	23	extinct, extinction, endangered	0.193	0.180	0.161	0.212	0.254
	123	die, death, dying, dead	0.164	0.182	0.185	0.215	0.253
	95	inequality, welfare, poverty, homelessness	0.182	0.170	0.173	0.231	0.244
	10	pesticide, GMO, Monsanto	0.150	0.173	0.182	0.255	0.240
	106	citizen, citizenship	0.152	0.174	0.218	0.216	0.240
	87	religion, religious, god	0.196	0.177	0.163	0.226	0.239
LOW	26	black hole, big bang, singularity	0.236	0.234	0.189	0.182	0.159
	17	dinosaur	0.225	0.237	0.198	0.183	0.156
	72	marijuana, pot, cannabis, THC, CBD	0.233	0.196	0.237	0.180	0.154
	55	stem cell	0.249	0.229	0.187	0.182	0.154
	37	genetic, DNA, gene, mutation	0.221	0.233	0.241	0.160	0.146
	36	cell, molecule, protein, enzyme	0.250	0.250	0.196	0.163	0.140

CHAPTER 6
CONCLUSION

6.1 Summary

In this thesis, we presented several new approaches toward the analysis and utilization of click-based reactions. We refined this approach through the analysis of several datasets of Facebook shares of scholarly articles posted on public pages of science and used this data to explore the emotional responses users have to scientific topics. We began in Chapter 3 with an study of the public pages onto which our articles were shared. We discovered some patterns of information sharing in our data that helped prepare for further analyses of the posts in our data. This preliminary study also gave us a better understanding of the data through a close examination of a small sample of posts. In Chapter 4, we suggested a method of transforming click-based reactions for analysis that was modeled after the TF-IDF statistic. We drew on concepts from behavioral psychology to develop several metrics for measuring aggregate user behavior with these transformed features. Finally, we used LDA topic modeling and statistical testing to find surprising emotional responses to scientific topics.

In the final study of this thesis, which we presented in Chapter 5, we used a semi-supervised approach toward LDA topic modeling to discover more granular topics of science in our data. We then used this LDA model to represent our documents as mixtures of scientific topics. These mixtures were then used as input to logistic regression models, which predicted the likelihood of seeing each click-based reaction in the presence of our topics of science. These logit models gave us the ability to make inferences about the relationship between our topics and the reactions they are associated with. This study took us from having emotional reactions to articles to having emotional responses to topics of science—a move toward generalization that we wanted from the outset of this thesis.

Finding instances of emotional diversity or strong intensity can be beneficial in understanding community dynamics. Conflict and division can be the fault lines on which groups divide, and discovering ways to identify them and make interventions if necessary can improve member cohesion. These lines of division also allow us to appreciate the diversity that exists on many online platforms.

6.2 Future Work

Following the contributions of this thesis, we plan on using social network analysis to gain better knowledge of how emotions spread within scientific communities. We will investigate the ways emotions affect the dispersion of disinformation and misinformation online, and the role they play in the effective communication of findings to non-experts. Exploring these and other related questions will eventually lead to better outcomes for research and will improve our understanding of the influence emotions have in shaping communication between scientists and the public.

Exploring how emotional response patterns change over time through a longitudinal study would also be a possible application of some of the methods and metrics we have developed in this thesis. Gaining a better understanding about how emotional responses to topics of science, or even to specific scientific articles, would give scientists and science communicators a better understanding of the changing needs in their fields, and urge them to develop better practices in a fast-moving world where consensus forms and dissipates rapidly.

One further aspect of the logistic models produced in Chapter 5 is that we can use them to predict what the response in click-based reactions will be for new documents. Predicting reactions for documents involves first encoding them as topic compositions with the same LDA model we used to train the logit model, and then feeding those topic distributions into our CMLogit class that has been trained with our data. Our CMLogit class has a method denoted by *predict* that takes a new vector or matrix $\mathbf{X} \in \mathbb{R}^D$ and returns an probability distribution of reaction types that the document or documents are estimated to receive. This is an interesting way to use our model, and could have some positive outcomes for researchers or science communicators who want to predict the response that a given post will receive. The use of our model for predictive tasks is outside the scope of this thesis however; before being assured that these predictions are reliable to a specified degree of certainty, more work would need to be done to gauge the accuracy of these predictions when applied to unseen test cases (i.e., we would need to compare our predicted reaction distributions to the known responses those posts received for documents that were not in our training data).

REFERENCES

[1] Dan Noyes. The top 20 valuable facebook statistics, June 2019. https://zephoria.com/top-15-valuable-facebook-statistics/, (visited on March 9,2020).

[2] Rodrigo Costas, Zohreh Zahedi, and Paul Wouters. Do "altmetrics" correlate with citations? extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, 66(10):2003–2019, 2015.

[3] Cassidy R. Sugimoto, Sam Work, Vincent Larivière, and Stefanie Haustein. Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology*, 68(9):2037–2062, 2017.

[4] Mike Thelwall and Tamara Nevill. Could scientists use altmetric.com scores to predict longer term citation counts? *Journal of Informetrics*, 12(1):237 – 248, 2018.

[5] Fereshteh Didegah, Timothy D. Bowman, and Kim Holmberg. On the differences between citations and altmetrics: An investigation of factors driving altmetrics versus citations for finnish articles. *Journal of the Association for Information Science and Technology*, 69(6):832–843, 2018.

[6] Hamed Alhoori and Richard Furuta. Do altmetrics follow the crowd or does the crowd follow altmetrics? In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, pages 375–378, 2014.

[7] Hamed Alhoori, Mohammed Samaka, Richard Furuta, and Edward A Fox. Anatomy of scholarly information behavior patterns in the wake of academic social media platforms. *International Journal on Digital Libraries*, 20(4):369–389, 2019.

[8] Asura Enkhbayar, Stefanie Haustein, Germana Barata, and Juan Pablo Alperin. How much research shared on facebook is hidden from public view? a comparison of public and private online activity around plos one papers, 2019.

[9] Mojisola Erdt, Aarthy Nagarajan, Sei-Ching Joanna Sin, and Yin-Leng Theng. Altmetrics: an analysis of the state-of-the-art in measuring research impact on social media. *Scientometrics*, 109(2):1117–1166, Nov 2016.

[10] Zohreh Zahedi and Rodrigo Costas. General discussion of data quality challenges in social media metrics: Extensive comparison of four major altmetric data aggregators. *PLOS ONE*, 13(5):1–27, 05 2018.

[11] Kimberley Collins, David Shiffman, and Jenny Rock. How are scientists using social media in the workplace? *PloS one*, 11(10), 2016.

[12] Juan Pablo Alperin. Geographic variation in social media metrics: an analysis of latin american journal articles. *Aslib Journal of Information Management*, 67(3):289, 2015.

[13] Heather C Lench, Sarah A Flores, and Shane W Bench. Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: a meta-analysis of experimental emotion elicitations. *Psychol. Bull.*, 137(5):834–855, September 2011.

[14] Barbara L Fredrickson. The role of positive emotions in positive psychology. the broaden-and-build theory of positive emotions. *Am. Psychol.*, 56(3):218–226, March 2001.

[15] Philippe Verduyn, Iven Van Mechelen, Francis Tuerlinckx, and Klaus Scherer. The relation between appraised mismatch and the duration of negative emotions: Evidence for universality. *Eur. J. Pers.*, 27(5):481–494, 2013.

[16] Rebecca J Compton. The interface between emotion and attention: A review of evidence from psychology and neuroscience. *Behav. Cogn. Neurosci. Rev.*, 2(2):115–129, June 2003.

[17] Nilam Ram, Denis Gerstorf, Ulman Lindenberger, and Jacqui Smith. Developmental change and intraindividual variability: relating cognitive aging to cognitive plasticity, cardiovascular lability, and emotional diversity. *Psychol. Aging*, 26(2):363–371, June 2011.

[18] Gordon H Bower and Joseph P Forgas. Mood and social memory. In Joseph P Forgas, editor, *Handbook of affect and social cognition , (pp*, volume 457, pages 95–120. Mahwah, NJ, US, Lawrence Erlbaum Associates Publishers, xviii, 2001.

[19] Philippe Verduyn, Iven Van Mechelen, and Evelien Frederix. Determinants of the shape of emotion intensity profiles. *Cognition and Emotion*, 26(8):1486–1495, 2012.

[20] Philippe Verduyn, Ellen Delvaux, Hermina Van Coillie, Francis Tuerlinckx, and Iven Van Mechelen. Predicting the duration of emotional experience: two experience sampling studies. *Emotion*, 9(1):83–91, February 2009.

[21] Linda J Levine and Robin S Edelstein. Emotion and memory narrowing: A review and goal-relevance approach. *Cognition and Emotion*, 23(5):833–875, 2009.

[22] Björn Ross, Tobias Potthoff, Tim A Majchrzak, Narayan Ranjan Chakraborty, Mehdi Ben Lazreg, and Stefan Stieglitz. The diffusion of crisis-related communication on social media: an empirical analysis of facebook reactions. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[23] Daniel T Gilbert, Elizabeth C Pinel, Timothy D Wilson, Stephen J Blumberg, and Thalia P Wheatley. Immune neglect: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, 75(3):617–638, 1998.

[24] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[25] Jeffrey T. Hancock, Kailyn Gee, Kevin Ciaccio, and Jennifer Mae-Hwah Lin. I'm sad you're sad: Emotional contagion in cmc. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, pages 295–298, New York, NY, USA, 2008. ACM.

[26] Afarin Pirzadeh and Mark S. Pfaff. How do you im when you get emotional? In *Proceedings of the 18th International Conference on Supporting Group Work*, GROUP '14, pages 243–249, New York, NY, USA, 2014. ACM.

[27] Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. Social clicks: What and who gets read on twitter? *SIGMETRICS Perform. Eval. Rev.*, 44(1):179–192, June 2016.

[28] Harish Varma Siravuri, Akhil Pandey Akella, Christian Bailey, and Hamed Alhoori. Using social media and scholarly text to predict public understanding of science. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, pages 385–386, New York, NY, USA, 2018. ACM.

[29] Felipe Taliar Giuntini, Larissa Pires Ruiz, Luziane De Fátima Kirchner, Denise Aparecida Passarelli, Maria De Jesus Dutra Dos Reis, Andrew Thomas Campbell, and Jó Ueyama. How do i feel? identifying emotional expressions on facebook reactions using clustering mechanism. *IEEE Access*, 7:53909–53921, 2019.

[30] Lisa Graziani, Stefano Melacci, and Marco Gori. Jointly learning to detect emotions and predict facebook reactions. In Igor V. Tetko, Věra Kůrková, Pavel Karpov, and Fabian Theis, editors, *Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series*, pages 185–197, Cham, 2019. Springer International Publishing.

[31] Bin Tareaf Raad, Berger Philipp, Hennig Patrick, and Meinel Christoph. Aseds: Towards automatic social emotion detection system using facebook reactions. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pages 860–866, June 2018.

[32] Ismail Badache and Mohand Boughanem. Emotional social signals for search ranking. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1053–1056, New York, NY, USA, 2017. ACM.

[33] Cole Freeman, Mrinal Kanti Roy, Michele Fattoruso, and Hamed Alhoori. Shared feelings: Understanding facebook reactions to scholarly articles. In *Proceedings of the 18th Joint Conference on Digital Libraries*, JCDL '19, page 301–304. IEEE Press, 2019.

[34] Cole Freeman, Hamed Alhoori, and Murtuza Shahzad. Measuring the diversity of facebook reactions to research. *Proc. ACM Hum.-Comput. Interact.*, 4(GROUP), January 2020.

[35] Ye Tian, Thiago Galery, Giulio Dulcinati, Emilia Molimpakis, and Chao Sun. Facebook sentiment: Reactions and emojis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 11–16, 2017.

[36] Florian Krebs, Bruno Lubascher, Tobias Moers, Pieter Schaap, and Gerasimos Spanakis. Social emotion mining techniques for facebook posts reaction prediction. *CoRR*, abs/1712.03249, 2017.

[37] Angelo Basile, Tommaso Caselli, and Malvina Nissim, editors. *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017), Rome, Italy, December 11-13, 2017*, volume 2006 of *CEUR Workshop Proceedings*, 2017.

[38] Markus Rohde, Leonard Reinecke, Bernd Pape, and Monique Janneck. Community-building with web-based systems – investigating a hybrid community of students. *Computer Supported Cooperative Work (CSCW)*, 13(5):471–499, Dec 2004.

[39] Anne Hewitt and Andrea Forte. Crossing boundaries: Identity management and student/faculty relationships on the facebook. *Poster presented at CSCW, Banff, Alberta*, pages 1–2, 2006.

[40] Moira Burke and Mike Develin. Once more with feeling: Supportive responses to social sharing on facebook. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW '16, pages 1462–1474, New York, NY, USA, 2016. ACM.

[41] Paul Thagard and Fred W Kroon. Emotional consensus in group decision making. *Mind & Society*, 5(1):85–104, 2006.

[42] Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 933–943, 2018.

[43] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social Media Mining: An Introduction*. Cambridge University Press, 2014.

[44] J Niels Rosenquist, James H Fowler, and Nicholas A Christakis. Social network determinants of depression. *Molecular psychiatry*, 16(3):273, 2011.

[45] Syed S Mahmood, Daniel Levy, Ramachandran S Vasan, and Thomas J Wang. The framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet*, 383(9921):999 – 1008, 2014.

[46] Rui Fan, Jichang Zhao, Yan Chen, and Ke Xu. Anger is more influential than joy: sentiment correlation in weibo. *PLoS One*, 9(10):e110184, October 2014.

[47] Vimala Balakrishnan, Vithyatheri Govindan, Noreen Izza Arshad, Liyana Shuib, and Ernest Cachia. Facebook user reactions and emotion: An analysis of their relationships among the online diabetes community. *Malaysian Journal of Computer Science*, pages 87–97, 2019.

[48] Pete Burnap, Rachel Gibson, Luke Sloan, Rosalynd Southern, and Matthew Williams. 140 characters to victory?: Using twitter to predict the UK 2015 general election. *Elect. Stud.*, 41:230–233, March 2016.

[49] Tapio Vepsäläinen, Hongxiu Li, and Reima Suomi. Facebook likes and public opinion: Predicting the 2015 finnish parliamentary elections. *Gov. Inf. Q.*, 34(3):524–532, September 2017.

[50] Lutz Bornmann. Validity of altmetrics data for measuring societal impact: A study using data from altmetric and f1000prime. *Journal of Informetrics*, 8(4):935–950, 2014.

[51] Hamed Alhoori, Richard Furuta, Myrna Tabet, Mohammed Samaka, and Edward A Fox. Altmetrics for country-level research assessment. In *International Conference on Asian Digital Libraries*, pages 59–64. Springer, 2014.

[52] Hamed Alhoori, Sagnik Ray Choudhury, Tarek Kanan, Edward Fox, Richard Furuta, and C Lee Giles. On the relationship between open access and altmetrics. In *Proceedings of the iConference*, 2015.

[53] Hamed Alhoori and Richard Furuta. Recommendation of scholarly venues based on dynamic user interests. *Journal of Informetrics*, 11(2):553–563, 2017.

[54] Hamed Alhoori. How to identify specialized research communities related to a researcher's changing interests. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 239–240. ACM, 2016.

[55] Harish Varma Siravuri and Hamed Alhoori. What makes a research article newsworthy? *Proceedings of the Association for Information Science and Technology*, 54(1):802–803, 2017.

[56] Bharat Kale, Harish Varma Siravuri, Hamed Alhoori, and Michael E Papka. Predicting research that will be cited in policy documents. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 389–390. ACM, 2017.

[57] Christian Bailey, Bharat Kale, Jamieson Walker, Harish Varma Siravuri, Hamed Alhoori, and Michael E Papka. Exploring features for predicting policy citations. In *Proceedings of the 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–2. IEEE, 2017.

[58] Abdul Rahman Shaikh and Hamed Alhoori. Predicting patent citations to measure economic impact of scholarly research. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 400–401. IEEE, 2019.

[59] Sammi Krug. Reactions now available globally, February 2016. URL: `https://newsroom.fb.com/news/2016/02/reactions-now-available-globally/` (visited on Nov. 30, 2018).

[60] Pritam Shah. Facebook's new reactions are being used more - a lot more., June 2018. URL: `https://www.quintly.com/blog/new-facebook-reaction-study` (visited on Jan. 24, 2019).

[61] Marcel Proust. *Remembrance of things past; translated by C.K. Scott Moncrieff and Terence Kilmartin.* Random House New York, 1981.

[62] Skeptical science. URL: `https://www.facebook.com/pg/SkepticalScience/` (visited on March 9, 2020).

[63] The conversation. URL: `https://www.facebook.com/ConversationEDU/` (visited on March 9, 2020).

[64] Who we are. URL: `https://theconversation.com/us/who-we-are` (visited on March 9, 2020).

[65] Food science and nutrition. URL: `https://www.facebook.com/elsevierfoodscience/` (visited on March 9, 2020).

[66] Chemistry world. URL: https://www.facebook.com/ChemistryWorld/ (visited on March 9, 2020).

[67] American society for microbiology. URL: https://www.facebook.com/asmfan/ (visited on March 9, 2020).

[68] Nobel prize. URL: https://www.facebook.com/nobelprize/ (visited on March 9, 2020).

[69] The new england journal of medicine. URL: https://www.facebook.com/TheNewEnglandJournalofMedicine/ (visited on March 9, 2020).

[70] National stop harper campaign. URL: https://www.facebook.com/NationalStopHarperCampaign/ (visited on March 9, 2020).

[71] Alt us national park service. URL: https://www.facebook.com/AltUSNationalParkService/ (visited on March 9, 2020).

[72] Rethink the link. URL: https://www.facebook.com/rethinkthelink/ (visited on March 9, 2020).

[73] Australian family association. URL: https://www.facebook.com/ausfamilyassociation/ (visited on March 9, 2020).

[74] Guido Palazzo and Kunal Basu. The ethical backlash of corporate branding. *Journal of Business Ethics*, 73(4):333–346, Jul 2007.

[75] Threatened species commissioner. URL: https://www.facebook.com/TSCommissioner/ (visited on March 9, 2020).

[76] J.C.Z. Woinarski, B.P. Murphy, S.M. Legge, S.T. Garnett, M.J. Lawes, S. Comer, C.R. Dickman, T.S. Doherty, G. Edwards, A. Nankivell, D. Paton, R. Palmer, and L.A. Woolley. How many birds are killed by cats in australia? *Biological Conservation*, 214:76 – 87, 2017.

[77] Mike bloomberg. URL: `https://www.facebook.com/mikebloomberg/` (visited on March 9, 2020).

[78] Personal trainer luca gorgoglione. URL: `https://www.facebook.com/vidoveterompereilculo/` (visited on March 9, 2020).

[79] Lindo bacon community. URL: `https://www.facebook.com/LindoBaconX/` (visited on March 9, 2020).

[80] Menno henselmans. URL: `https://www.facebook.com/MennoHenselmans/` (visited on March 9, 2020).

[81] Quackwatch. URL: `https://www.facebook.com/Quackwatch/` (visited on March 9, 2020).

[82] National universal medicare for all. URL: `https://www.facebook.com/M4ANOW2020/` (visited on March 9, 2020).

[83] Dr. bogner. URL: `https://www.facebook.com/drbogner/` (visited on March 9, 2020).

[84] The commonwealth fund. URL: `https://www.facebook.com/commonwealthfund/` (visited on March 9, 2020).

[85] arxiv-social. URL: `https://www.facebook.com/arxivsocial/` (visited on March 9, 2020).

[86] arxiv sanity. URL: https://www.facebook.com/ArxivSanity/ (visited on March 9, 2020).

[87] arxiv general relativity and quantum cosmology. URL: https://www.facebook.com/arxivgrgc/ (visited on March 9, 2020).

[88] arxiv. URL: https://www.facebook.com/ARXIV-1658417297558124/ (visited on March 9, 2020).

[89] Machine learning research at arxiv. URL: https://www.facebook.com/arxivML/ (visited on March 9, 2020).

[90] Wolf conservation center. URL: https://www.facebook.com/nywolforg/ (visited on March 9, 2020).

[91] News from science. URL: https://www.facebook.com/ScienceNOW/ (visited on March 9, 2020).

[92] Science magazine. URL: https://www.facebook.com/ScienceMagazine/ (visited on March 9, 2020).

[93] The matter. URL: https://www.facebook.com/thematterco/ (visited on March 9, 2020).

[94] Lebanese researchers. URL: https://www.facebook.com/lebaneseresearchers/ (visited on March 9, 2020).

[95] Exposing feminism 2.0. URL: https://www.facebook.com/ExposingFeminism2/ (visited on March 9, 2020).

[96] Whitney Stark. Assembled Bodies: Reconfiguring Quantum Identities. *the minnesota review*, 2017(88):69–82, 05 2017.

[97] World bank publications. URL: `https://www.facebook.com/worldbankpublications/` (visited on March 9, 2020).

[98] Erin M. Sumner, Rebecca A. Hayes, Caleb T. Carr, and Donghee Yvette Wohn. Assessing the cognitive and communicative properties of facebook reactions and likes as lightweight feedback cues. *First Monday*, 25(2), Jan. 2020.

[99] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

[100] Natalya N. Bazarova, Yoon Hyung Choi, Victoria Schwanda Sosik, Dan Cosley, and Janis Whitlock. Social sharing of emotions on facebook: Channel differences, satisfaction, and replies. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 154–164, New York, NY, USA, 2015. ACM.

[101] John C. Besley and Matthew Nisbet. How scientists view the public, the media and the political process. *Public Understanding of Science*, 22(6):644–659, 2013.

[102] David van Dijk, Ohad Manor, and Lucas B. Carey. Publication metrics and success on the academic job market. *Current Biology*, 24(11):R516 – R517, 2014.

[103] Qing Ke, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, 112(24):7426–7431, 2015.

[104] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R.* Springer Publishing Company, Incorporated, 2014.

[105] Andrew Y Ng and Michael I Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848, 2002.

[106] Jon D Mcauliffe and David M Blei. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128, 2008.

[107] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213. Association for Computational Linguistics, 2012.

[108] Shachi H Kumar, Eda Okur, Saurav Sahay, Juan Jose Alvarado Leanos, Jonathan Huang, and Lama Nachman. Context, attention and audio feature explorations for audio visual scene-aware dialog. *arXiv preprint arXiv:1812.08407*, 2018.

[109] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

[110] Tim Weninger, Yonatan Bisk, and Jiawei Han. Document-topic hierarchies from document graphs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 635–644, 2012.

[111] B. Peng, B. Zhang, L. Chen, M. Avram, R. Henschel, C. Stewart, S. Zhu, E. Mccallum, L. Smith, T. Zahniser, J. Omer, and J. Qiu. Harplda+: Optimizing latent dirichlet allocation for parallel efficiency. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 243–252, Dec 2017.

[112] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[113] Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. Social Clicks: What and Who Gets Read on Twitter? In *ACM SIGMETRICS / IFIP Performance 2016*, Antibes Juan-les-Pins, France, June 2016.

[114] Jafar Alqatawna, Hossam Faris, Khalid Jaradat, Malek Al-Zewairi, Omar Adwan, et al. Improving knowledge based spam detection methods: The effect of malicious related features in imbalance data distribution. *International Journal of Communications, Network and System Sciences*, 8(05):118, 2015.

[115] Shigang Liu, Yu Wang, Jun Zhang, Chao Chen, and Yang Xiang. Addressing the class imbalance problem in twitter spam detection using ensemble learning. *Computers & Security*, 69:35–49, 2017.

[116] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[117] Shiven Sharma, Colin Bellinger, Bartosz Krawczyk, Osmar Zaiane, and Nathalie Japkowicz. Synthetic oversampling with the majority class: A new perspective on handling extreme imbalance. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 447–456. IEEE, 2018.

[118] J Aitchison. The statistical analysis of compositional data. *J. R. Stat. Soc. Series B Stat. Methodol.*, 44(2):139–160, January 1982.

[119] J J Egozcue, V Pawlowsky-Glahn, G Mateu-Figueras, and C Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Math. Geol.*, 35(3):279–300, April 2003.

[120] K Hron, P Filzmoser, and K Thompson. Linear regression with compositional explanatory variables. *J. Appl. Stat.*, 39(5):1115–1128, May 2012.

[121] Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. *Modeling and Analysis of Compositional Data*. John Wiley & Sons, February 2015.

[122] John Aitchison. The statistical analysis of geochemical compositions. *Journal of the International Association for Mathematical Geology*, 16(6):531–564, August 1984.

[123] Wes McKinney. Data structures for statistical computing in python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.

[124] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[125] Dong-Hui Li and Masao Fukushima. On the global convergence of the bfgs method for nonconvex unconstrained optimization problems. *SIAM Journal on Optimization*, 11(4):1054–1064, 2001.

[126] Roger Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.

[127] John E Dennis Jr and Robert B Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*, volume 16. Siam, 1996.

[128] PEARSON and K. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Royal Soc., London, Proc.*, 60:489–502, 1897.

[129] J Aitchison, C Barceló-Vidal, J A Martín-Fernández, and V Pawlowsky-Glahn. Logratio analysis and compositional distance. *Math. Geol.*, 32(3):271–275, April 2000.

APPENDIX A

COMPOSITIONAL DATA AND REGRESSION

This appendix provides a brief explanation of the problem of multicollinearity and compositional data in statistical modeling and machine learning. This discussion is set apart from the main body of the thesis because it is more technical in nature and focuses on issues of mathematics rather than on applications. The ideas are discussed here because they motivate many of the choices we made in Chapter 5 in building our models with LDA data (which is compositional), and because they provide a broader context for some of our analysis there. Though this appendix provides an overview of the mathematics of compositions, it also provides some examples of applications. Toward this latter goal, an example of the challenges we face when modeling with compositional data that uses linear regression. We felt that linear regression better highlights the problems and allows for better geometric demonstrations than does logistic regression—though the problems is virtually the same in both cases. The final section, which is devoted to discussing logratio transformations of compositional data, can be used as further clarification of the techniques used in Chapters 5 to solve our multicollinearity problem.

A.1 Compositional Data and the Multicollinearity Problem

Several studies have used topic models as a way to generate quantifiable features from text. These new features are often used in model building for predictive or explanatory tasks; but there are some thorny issues with using topic mixtures in this way. In supervised learning tasks, data is usually divided into independent variables (sometimes called explanatory data, predictors, or regressors in the case of regression) and dependent variables (otherwise known as response or target variables). Patterns of change in the independent variables are tied to variations in the dependent variables, allowing modelers to predict responses to future stimuli of the same form. The resulting model is analyzed based on its performance and what it can tell us about the discovered patterns in the data.

An assumption of most model types is that the explanatory variables are all independent. Problems in multivariate data analysis arise when there is multicollinearity, or strong correlations between multiple independent variables. When one feature can be derived within a small margin of error from another, the two variables do not meet this requirement of independence. They carry the same information and form what is known as a data singularity. If these multicollinear features are used to train a model, the results will more than likely be biased in one way or another and lead to bad inferences on the part of the modeler. As such, it is common practice in modeling to check data for multicollinearity and root it out in an appropriate manner, either by removing one of each correlated pair or by applying an appropriate transformation to your data.

One type of multicollinearity is when explanatory variables contain relative rather than absolute information. This unique type of singularity is known as compositional data, in which multiple components describe the parts of a whole. Meaningful information in such cases is carried by the ratio of the parts to the whole and not by individual parts themselves. Compositional data can take the form of proportions, frequencies, or probabilities, in which all components sum to 1, percentages, in which components all sum to 100, or concentrations, in which some other arbitrary scale such as ppm (parts per million) or ppb (parts per billion) is used to report parts of a composition. As early as 1897 Karl Pearson noted that researchers should be wary of making statistical inferences from such data in his paper On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs [128]. Despite early knowledge of the problem researchers have often given in to the temptation to ignore the special qualities of compositional data in their analyses. The Scottish statistician John Aitchison devoted much of his career to exploring compositional data and developing methods for applying traditional statistical methods to this special type of data [118, 122, 129].

Since the early 2000s there has been an effort to fully define the special problems pertaining to compositional data, and to explore the methods that can produce useful information from it [119, 121, 120].

The multicollinearity problem lies in the fact that though there may be $d \geq 2$ components in a composition, there are only $(d-1)$ independent values. The final dimension d is fully determined by the other components. The other d-1 components are still correlated: when one goes up, the cumulative value of the others necessarily goes down by that same amount. This type of multicollinearity is especially insidious as the correlation coefficients of the components are not always high enough to raise warning flags. Only a deep understanding of the data we are dealing with will give us insight into its special properties and point our analysis in the appropriate direction.

The sample space of compositional data is not real space, but rather a constrained subset of it known as the simplex. This unique geometric space takes the form of a generalized triangle of arbitrary dimensions: a one-simplex is a line, a two-simplex a triangle, a three-simplex a tetrahedron, and on and on into higher dimensions. The sample space of compositions is defined generally as a set of features S where:

$$S^D = \{x = [x_1, x_2, ..., x_D] \in \mathbb{R}^D | x_i \geq 0, i = 1, 2, ..., D; \sum_{i=1}^{D} x_i = k\} \tag{A.1}$$

where D is the dimensionality of the data and k is an arbitrary positive constant.

A.1.1 Closure Operation

Though k may be arbitrary, we might want to normalize our data to facilitate comparison with other compositions or for greater comprehensibility of the values. Normalization of a composition is known as closure. This operation involves changing the constant value k which all vectors sum to. It is common to set $k = 1$ for compositions, as this makes comparison easier by the direct analogy to proportions. For a set of D components $\mathbf{x} = [x_1, x_2, ..., x_D]\mathbb{R}^D$ and $x_i \geq 0$ for all $i = 1, 2, ..., D$, closure is achieved by the following procedure:

$$C(x, k) = [\frac{kx_1}{\sum_{i=1}^{D} x_i}, \frac{kx_2}{\sum_{i=1}^{D} x_i}, ..., \frac{kx_D}{\sum_{i=1}^{D} x_i}] \tag{A.2}$$

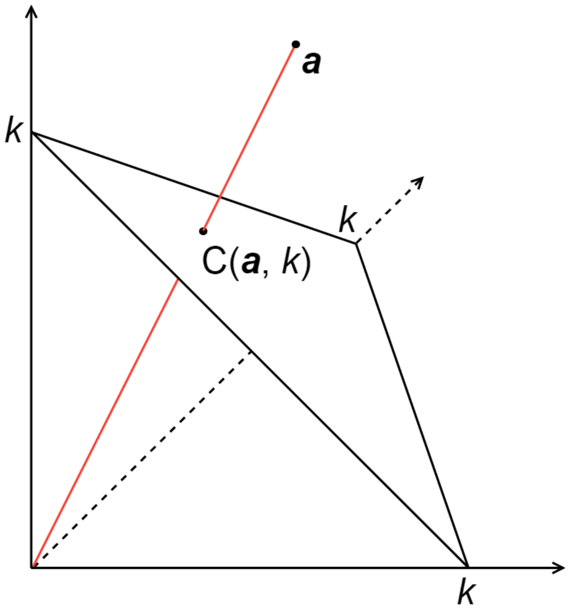


Figure A.1: Representation of the closure operation in a three-dimensional space. $C(\mathbf{a}, k)$ and $\mathbf{a}$ lie on a vector that passes from the origin into the positive orthant. $C(\mathbf{a}, k)$ represents closure of $\mathbf{a}$ into a composition whose parts sum to k. It represents a projection of $\mathbf{a}$ into a simplex of different proportions. Closure is a scale-invariant operation, in that it equivalently transforms all parts of a composition and maintains their proportionality.

where $k > 0$. Two vectors x and y, which have different constant sums k, are compositionally equivalent if $C(x, k) = C(y, k)$. Figure A.1 demonstrates the geometry of the closure operation. Points $\mathbf{a}$ and $C(\mathbf{a}, k)$ both lie on a vector that passes from the origin into the positive orthant. $C(\mathbf{a}, k)$ brings the three-part composition that represents the coordinates of $\mathbf{a}$ into a 2-simplex where all parts sum to k. We should note that $\mathbf{a}$ is also a composition in a simplex with a higher value of k, and also that $\mathbf{a}$ is equivalent to $C(\mathbf{a}, k)$ by the definition of closure. The difference between $\mathbf{a}$ and $C(\mathbf{a}, k)$ is the scale at which the two points are shown. By the definition of a composition, the magnitude of the vector that passes through $\mathbf{a}$ and $C(\mathbf{a}, k)$ carries no important information: it is only the proportion of the parts that has meaning here.

A.1.2 Scale Invariance

Figure A.1 also hints at one of the most important conditions we require of operations on compositional data: all transformations of compositions should be scale invariant. As defined in Aitchison [2] and again in Pawlowsky-Glahn, Egozcue, et al. [6], scale invariance ensures that any transformation process applied to two compositionally equivalent points will yield compositionally equivalent points. Expressed mathematically, we state this property as follows:

$$C(f(\lambda x), k) = C(f(x), k) \tag{A.3}$$

where f is a continuous function in $\mathbb{R}^D$ throughout the positive orthant and $\lambda \in \mathbb{R}_+$. Scale invariance is achieved when f is a 0-degree homogeneous function that affects all parts of composition $\mathbf{x}$ equivalently. Scale invariance is a property of compositions, which consist of dimensionless quantities. Thus any transformation of a composition must preserve this original and necessary property.

A.1.3 Subcompositions

In practice compositions can be made up of a great many parts, not all of which are interesting objects for study. Often a researcher may only want to focus on a specific subset of a composition. Even though this subset does not represent the totality of elements present, the researcher may decide for one reason or another to treat it as such. This subset of elements is called a *subcomposition*. Examples of the importance of subcompositions abound but are frequently found in the geological sciences, especially in geology where a main concern is the makeup of rocks and how they change over time. When a geologist studies a type of rock they usually focus on the primary components of the material: accounting for every possible element present is difficult and would not significantly affect the findings. Geologists tend to identify and focus on important subcompositions of all the materials in a sample. We define a subcomposition as a selected portion of an overall composition that is obtained by applying the closure operation to a subvector $\mathbf{X}_S$ of a composition $\mathbf{X}$, where S is a set of indices taken from $\mathbf{X}$.

We may also wish to combine several elements of a composition and treat them as a single component. This may be done for a number of reasons: we may see a similarity between elements and wish to them as a unit; we may want to partition our composition into different subgroups, whose proportional relationships we are interested in studying; we may also be interested in isolating specific elements by partitioning all other elements into a single group. We call this process of combining components amalgamation. We define amalgamation as:

$$\mathbf{X}_A = \sum_{i \in A} x_i \tag{A.4}$$

where $\mathbf{X}$ is a composition, A is a selection of indices from $\mathbf{X}$, and $\mathbf{X}_A$ is an amalgamated part or amalgamated component.

A.1.4 Example Composition

As an example, imagine we are making an unleavened bread with a recipe that calls for 2 cups flour, $1\frac{1}{3}$ cups water, $\frac{1}{3}$ cup olive oil, and a pinch of salt. Our recipe yields two loaves. Say we want to have a recipe that works for n loaves of bread, and decide to figure out the exact proportions we would need for any batch size. We could convert our recipe into a composition of ingredients. We do not quite know how to quantify a "pinch" of salt, so we choose to disregard that element in our recipe of proportions and only focus on the set of ingredients containing flour, water, and oil. We find that our recipe is $[Flour \approx 0.55, Water \approx 0.36, Oil \approx 0.09]$, where $k = 1$. Here we have proportions, which are dimensionless quantities. Our new recipe can be used to determine the proportions of ingredients for any number of loaves.

Now imagine we apply what we think are the correct proportions to batch of five loaves. We follow all the correct procedures for combining ingredients, kneading, and baking the bread, but the result is not quite right to our taste, and we want to know what we did wrong. We have a very advanced kitchen with scientific equipment. Among our appliances is a mass spectrometer that can tell us the compositional elements in samples of our bread. Assuming that our kneading process has distributed the ingredients throughout the dough evenly, we take a pinch of bread and put it through the mass spectrometer to learn what elements are there. We find that our ingredients $[Flour, Water, Oil]$ are present in the following proportions: $[0.62, 0.32, 0.06]$. So what went wrong here? It is impossible to say exactly where our bread-making process went awry, but there are several possibilities to consider:

1. We added the right amount of flour for our batchsize, but only used about 79% of the water we should have and roughly 60% of the correct amount of oil;

2. We used the proper amount of water for the number of loaves we were making, but put in about 25% too much flour and 25% less oil than there should be; or

3. We used the correct amount of oil for our batchsize, but used almost 70% too much flour and 33% too much water.

Without absolute knowledge about the specific amount of any ingredient, or absolute knowledge of the total amount of ingredients used we cannot tell how we went wrong.

Upon further reflection, we decide the flavor of the bread is not really the problem, but that the consistency of our product is off. In this case, it may be beneficial to analyze the proportion of the subcompositions of wet ingredients and dry ingredients. Using the subvector for dry ingredients as $[Flour]$, and that for wet ingredients $[Water, Oil]$, we amalgamate the values in our compositional recipe as $[dry \approx 0.55, wet \approx 0.45]$. We can do the same with the elements of the bread we made, resulting in the composition: $[dry \approx 0.62, wet \approx 0.38]$. Following the same logic as above, we are unable to determine exactly where we went wrong in our proportions. We can say for certain that the ratio ingredients is wrong, and that we have too great a proportion of dry ingredients to wet ingredients. We cannot, however, say whether we: (a) added up to 33% too much of our dry ingredients, which would result in a total yield that is about 18% increased over what we expected, (b) used up to 25% too much of our combined wet ingredients, which give a total yield a little more than 11% less than we hoped, or (c) whether we used some combination of (a) and (b).

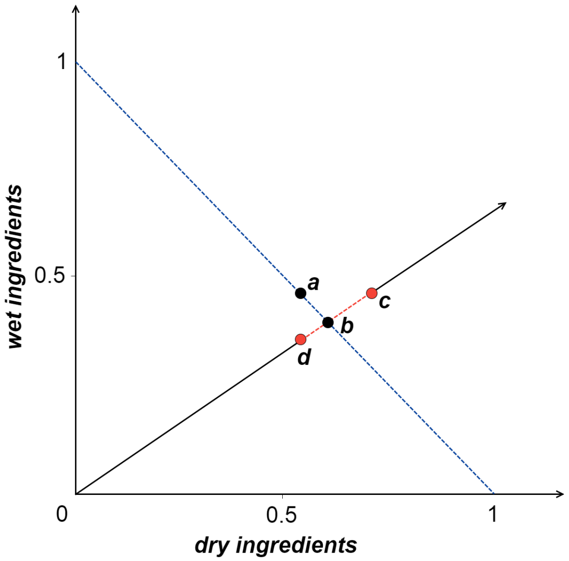


Figure A.2: A geometric representation of scale invariance. Given our example of our bread recipe, the proportion of dry ingredients to wet ingredients is shown. Point **a** is the correct proportions for our recipe, **c** represents the case where we have used too much of our dry ingredients but the correct amount of wet ingredients, **d** represents the case where we have used the proper amount of dry ingredients but too few wet ingredients, and **b** represents the closure operations $C(\mathbf{c}, 1)$ and $C(\mathbf{d}, 1)$. Notice that **b**, **c**, and **d** all fall on a vector passing through the origin. The three points **b**, **c**, and **d** have different values of k.

Figure A.2 gives a geometric representation of our mistake in proportions of wet to dry ingredients and the possible ways that we went wrong. Since we cannot say with certainty where our mistake was, we can only outline the possible ways in which we went wrong. Point **a** is the proportions of ingredients prescribed by our recipe. Points **c** and **d** are possible locations of our actual proportions, and point **b** is the closure of **c** and **d**.

Our bread-making example demonstrates many of the principles and operations of compositional data that we have outlined. By eliminating salt from consideration, we focused on the subcomposition containing flour, water, and olive oil. In combining the parts of water and oil into a single component representing wet ingredients, we used the process of amalgamation. Figure A.2 demonstrates the principle of scale invariance, showing how different mistakes of adding too much or too little of each component can result in equivalent compositions.

In addition to scale invariance There is one final condition of compositional data that should be met by any statistical analysis: *permutation invariance*

A.2 Linear Regression with Compositional Data

The singular nature of compositional data makes its use as a predictor problematic. Singular data confuses the results of linear regression models and can not be used for in such cases without additional constraints being placed on the model or without transforming the input data into a more meaningful form. The problem with using this type of data to fit regression models is most clearly demonstrated with a brief explanation of the geometry of regression, along with an example of building a model with absolute features and with compositional data.

Simple linear regression is a method of approximating a quantitative response variable Y with a single explanatory variable X. The assumption of this method is that there is a relationship between the two variables that takes can be expressed with a linear function. The relationship is described in the following form:

$$Y = \beta_0 + \beta_1 X \tag{A.5}$$

The two additional terms β_0 and β_1 demonstrate how values of X can be transformed into values of Y. Equation A.5 describes a line with the intercept term β_0 representing where the line crosses the y-axis and β_1 (known as the regression coefficient) describing the slope of the line. There are a number of methods for estimating these terms, the most common of which is the ordinary least squares method (OLS). OLS estimates the intercept and coefficient terms through the process of minimizing the residual sum of squares (RSS), which is defined by the following equation:

$$RSS = \sum_{i=1}^{n} (y - \hat{y})^2 \tag{A.6}$$

where $\hat{y}_i$ is the estimated value of y_i. $(y_i - \hat{y}_i)$ is known as the *residual*, or ε (error term), for an observation. It represents the difference between a value predicted by your model and the true value in your data. Another way of expressing error terms is as a value that can be added to Equation A.5 that yields the true value of the response:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{A.7}$$

Many of the inferences we are able to make from linear regression come from analysis of the intercept, the regression coefficient, and the residuals produced by our model. They allow us to gauge the effect that the explanatory variable has on the response. They let us measure our confidence in the relationship. They also allow us to see how well our model predicts the actual response values in our data.

In most cases where we want to use regression, we have multiple explanatory variables. This is the case in our marketing example. The relationship between our set of independent variables and response variable is expanded to include more coefficient terms:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_n x_{in} + \varepsilon_i \tag{A.8}$$

In this case, the effect that each explanatory variable has on the response is quantified by the coefficients $[1, ..., n]$.

A.2.1 Example: Marketing Data

Imagine we work for a marketing company and are trying to determine the extent to which spending money on TV and radio advertising will lead to an increase in sales.[1] We have data about previous times when the company spent money advertising on these media as well as the sales experienced right after those expenses. One approach we could take to analyze the relationship between expenses and sales would be to fit a regression model with our data. Such a method is powerful because it would allow us not only to predict future sales based on past observations, but also to analyze the role that each media plays in moving sales numbers and to see how spending on the different media interact with each other (known as *interaction effects*). Figure A.3 shows the OLS model fit to our data. OLS minimizes the errors across all three dimensions, yielding a plane of best fit that we can use to learn about the relationship between various expenditures and sales.

Now imagine this same problem where we have a fixed budget devoted to all media expenditures. Using more money on TV, for example, means we have less money to spend on radio. The nature of our problem has significantly changed with this simple constraint. Expenditures are no longer fully independent variables, as changes in one necessarily affect changes in the other. If we are still trying to fit a regression model, we must be aware of how our situation has changed. Figures A.4 and A.5 show the geometric situation of our new problem. In Figure A.4, a unit-sum 1-simplex is shown in positive 2-dimensional space with points representing our sales data from before as the proportion of TV and radio sales to total expenditures.

[1] The data in this section is taken from *Introduction to Statistical Learning*, by James et al. [104].

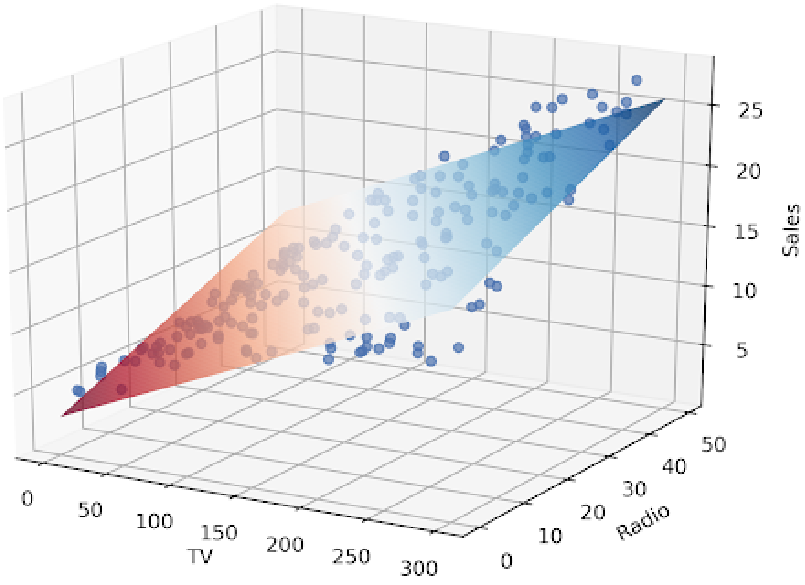


Figure A.3: An example of a regression model fitting data in a three-dimensional space.

In Figure A.5, that simplex is extended into 3-dimensions with the inclusion of the response y-axis, which represents the absolute value of sales, along which the values are free to vary. The values in this dataset will all lie within the two-dimensional plane marked in light blue in Figure A.5. An OLS model in this case would fit a plane that minimizes the errors between all points. Clearly, such a model would be next to meaningless as it would only be responding to points along a single vertical plane; the slope along one of the three dimensions can move in any direction and the fit of the model would not change.

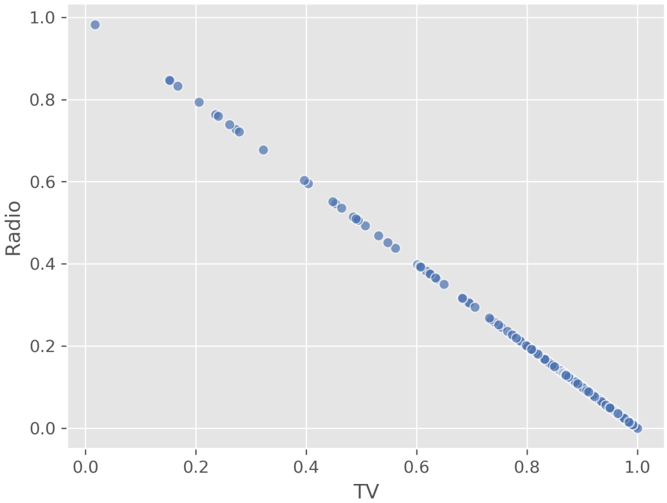


Figure A.4: A geometric representation of a 1-simplex (a k-simplex has $k+1$ vertices—here there are two vertices, one on each axis), which is a line of length $\sqrt{2}$. Two $(\mathbf{x}_1, \mathbf{x}_2)$ coordinates are plotted. Any points within this simplex must lie on the red line and will always sum to 1.

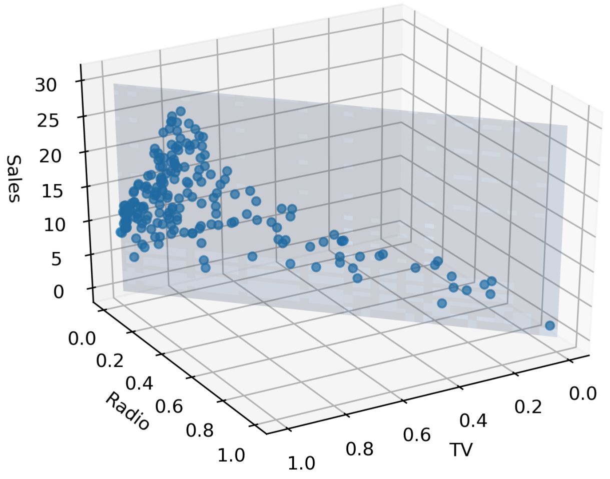


Figure A.5: The same 1-simplex depicted in Figure A.4, now with an added axis representing the absolute sales values.

APPENDIX B

COMPLETE LIST OF SEEDED TOPICS

The following is a list of the seeded topics we used for the LDA in Chapter 5:

0. environment, environmental, habitat, nature, natural
1. climate, weather, climatology, meteorology
2. climate change, global warming, climatechange
3. greenhouse gas, carbon dioxide, aerosol, greenhouse, ozone
4. carbon, carbon emission, emission
5. fossil fuel
6. ice, melt, ice cap, polar, iceberg
7. conservation, wildlife, conserve, conserved
8. fracking, frack
9. mine, mining
10. pesticide, gmo, roundup, monsanto, insecticide
11. plant, flower, tree, forest, rainforest, garden, botanical, park, national park, wood, botanic garden, palm tree, seed, germinate
12. deforestation
13. animal, species, specie, genus
14. cat, bird, dog, wolf, horse, monkey, gorilla, mouse
15. snake, reptile, treefrog
16. fish, fishing, fishery

17. dinosaur

18. entomology, entomologist, bug, beetle, insect

19. bee, honeybee, honey bee

20. mosquito

21. euthanasia, euthanize

22. veterinarian, vet, veterinary

23. extinct, extinction, endangered, endanger

24. physics, physic, physicist

25. cosmology, cosmological

26. black hole, big bang, singularity

27. universe, solar system, cosmo, cosmic, galaxy

28. particle, higgs, proton, electron, collider, hadron

29. quantum

30. ray, solar, sun, neutron star, star

31. lunar, moon

32. earth, planet

33. astronaut, rocket, shuttle, orbit, orbital

34. biology, biochemistry, microbiology, biologist, bio, microbiologist

35. chemistry, chemist, chemical

36. cell, molecule, protein, enzyme, molecular

37. genetic, dna, gene, genome, mutation, evolution, geneticist, genetically, nucleotide

38. brain, neuron, dopamine, synapse, neurology, neurologist

39. geology, geological, geologist, geologic, sediment, sedimentary, rock, igneous

40. geography, geographic, geographer

41. math, mathematic, geometry, geometric, mathematician, mathematical, trigonometry, algebra

42. probability, statistical, stat, statistic, distribution

43. robot, robotic, machine, mechanical

44. automate, automation

45. engineer, engineering

46. ai, algorithm, computer, deep learning, neural network

47. health, healthy, medicine, medical

48. healthcare, insurance

49. doctor, dr, md, nurse, physician

50. hospital, patient, hospitalize, hospitalization

51. sick, sickness, illness, disease, infection, infect, epidemic, virus, viral, bacteria, bacterial

52. surgery, surgeon, surgical

53. cancer, tumor, tumour, carcinogen, carcinogenic, cancerous

54. hiv, tcell

55. stem cell

56. blood, bleed, bleeding

57. inhibitor, active, inhibit

58. cardiology, cardiac, heart

59. vaccine, vaccinate, vaccinated, vaccination, immunization, immunize, immunology

60. unvaccinated, anti vaxxer

61. autism, autistic

62. mental health, mental

63. psychiatry, psychiatric, psychiatrist

64. mental disorder, mental illness

65. dementia, alzheimer, memory

66. suicide, suicidal, anxiety, anxious, loneliness, distress, depression

67. diet, dietary, sugar, oil, lipid, canola, vegetable oil, carb, food, nutrition, nutritional, metabolism, metabolic

68. obese, obesity, weight, overweight

69. vegan, vegetarian

70. allergy, allergic

71. gluten, gluten free

72. marijuana, pot, cannabis, thc, cbd, cannabinoid

73. lsd, heroin, cocaine, meth, methamphetamine

74. cigarette, smoking, smoker, tobacco

75. addict, addiction, addictive

76. pharmacology, pharma, pharmacist, pharmaceutical, prescription, prescribe

77. opiate, opioid

78. art, artistic, artist, culture, cultural, painting, literature

79. music, musical, song, sing, musician

80. language, linguistic, linguist

81. philosophy, philosophical, philosopher, logic, logical

82. ethic, ethical

83. unethical

84. history, historical, archaeology, archaeologist, archaeological

85. feminism, feminist

86. psychologist, psychology, psychological

87. religion, religious, god

88. bible, christian, christ, biblical, preacher, sin

89. buddhist, enlightenment

90. muslim, islam

91. hindu, hinduism

92. economy, economics, economist

93. crisis, recession

94. budget, spending, fiscal, money, cash, finance, financial

95. inequality, welfare, poverty, homeless, homelessness, impoverished

96. tax

97. phd, doctoral, congratulation, doi, postdoc, academic, academy, academia, professor, lecture, lecturer, teacher, teach, university, school, college, graduate, graduation, student, grade

98. find, could, link, association, factor, role

99. system, structure, organize, order, method, process, approach, level, activity

100. follow, contain, single, potential, manage, management, alter, change

101. peer review, journal, article, paper, period, study, volume, insight, assessment, apply, performance, evaluate, progression, methodology

102. politics, political, politician, policy, politic, government, national, nation, state, governmental

103. identity

104. liberal, conservative, republican, democrat, libertarian, govern, progressive

105. senate, senator, congress, congressional, parliament, parliamentary

106. citizen, citizenship

107. immigrant, immigration, refugee, immigrate

108. corruption, corrupt

109. poll, polling, election, vote, voter, voting, elect

110. epa, united state, america, american, doe, europe

111. trump, donald trump, turnbull, theresa may, merkel, prime minister, president, presidential

112. brexit

113. aca, affordable care, obamacare

114. gun, firearm, pistol, bullet, shoot, shooting, shooter

115. war, nuclear war, bomb, bombing, nuke, military

116. fight, fighting, attack

117. copyright, lawyer, law, attorney, legal, legally, court, judge

118. illegal, illicit

119. police, policeman, policing, cop, law enforcement

120. prison, jail, prisoner, imprison

121. crime, criminal

122. killing, kill, killer, murder, homicide, massacre, slaughter, execute, execution

123. die, death, dying, dead, obituary

124. gender

125. male, man, men, masculine, masculinity, boy

126. female, woman, women, girl

127. marriage, marry, husband, wife, married

128. sex, sexual

129. reproduction, mate, reproductive, pregnant, pregnancy, fertility, birth, baby

130. lesbian, gay, homosexual, samesex, same sex

131. bisexual

132. transgender, trans

133. heterosexual

134. fetal, aborted, abortion, fetus, embryo, embryonic

135. condom, contraception

136. abstinence

137. porn, pornography

138. penis, erection, uterus, breast, genital, vagina, vaginal, testicular, nipple

139. masturbation

140. orgasm, orgasms

141. diversity, diverse

142. sexism, discrimination, racism, racist

143. (curse words)

144. feces, toilet, bathroom

145. dummy, stupid, idiot, fool, foolish, crazy, insanity, absurd, absurdity

146. horrible, horrific, horrify, horror

147. harass, harassment, abuse, abuser, abused, violate

148. violent, violence

149. pain, painful, hurt, harm, harmful

150. disaster, catastrophe, catastrophic, ruin, damage, damaging, destroy, destruction, destructive, tragedy, tragic

151. danger, dangerous, peril, perilous, threat, threaten, threatened, unsafe

152. torture, tortuous

153. disturb, upset, disturbing

154. bad

155. good, great, grand, epic, neat

156. safe, safety

157. happy, joy, love, excited, excite

158. amazed, amaze, amazement, amazing

159. funny, laugh, humor

160. anger, angry, mad, rage

161. sad, cry, dismay

162. fear, afraid, fearful, scared, scare, scary, alarming

163. hate, hater