The Thesis Committee for James Spencer Evans
certifies that this is the approved version of the following thesis:

Characterizing the Relationship in Social Media Between Language and Perspective on Science-Based Reasoning as Justification for Belief

APPROVED BY

SUPERVISING COMMITTEE:

Jason Baldridge, Supervisor

Katrin Erk

Characterizing the Relationship in Social Media Between Language and Perspective on Science-Based Reasoning as Justification for Belief

by

James Spencer Evans, B.A.

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Arts

The University of Texas at Austin

May 2014

I dedicate this thesis to my mother and father.

Acknowledgments

I would like to acknowledge Jason Baldridge, Katrin Erk, and Joey Frazee for their guidance.

James Spencer Evans

The University of Texas at Austin
May 2014

Abstract

Characterizing the Relationship in Social Media Between Language and Perspective on Science-Based Reasoning as Justification for Belief

James Spencer Evans, M.A.
The University of Texas at Austin, 2014

Supervisor: Jason Baldridge

Beliefs that are not the result of science-based interpretation of evidence (e.g., belief in ghosts or belief that prayer is effective) are extremely common. Science enthusiasts have expressed interest in automatic detection of non-science-based claims. This thesis intends to provide some first steps toward a solution, specifically aimed at detecting Twitter users who are likely or unlikely to take a science-based perspective on all topics. As part of this thesis, a set a Twitter users was labeled as being either "pro-science" (i.e. as having the view that beliefs are rational if and only if they are in accord with science-based reasoning) or "non-pro-science" (i.e.

as having the view that beliefs may be reasonable even if they are not in accord with science-based reasoning). Word frequency ratios relative to a neutral dataset, and a simple topic alignment technique, suggest considerable linguistic divergence between the pro-science and non-pro-science users. High accuracy logistic regression classification using linguistic features of users' recent tweets support that idea. Supervised classification experiments suggest that the pro-science and non-pro-science perspectives are not only detectable from linguistic features, but that they can be abstracted away from particular topics (i.e. that the pro-science and non-pro-science perspectives are not inherently topic-specific). Results from distantly supervised classification suggest that using easily acquired, weakly labeled data may be preferable to the much slower process of individually labeling data for some applications, despite the pronounced inferiority to the fully supervised approach in terms of accuracy. The best classifier obtained in this thesis has an accuracy of 93.9%.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis is an exploration of the relationship between the linguistic features that characterize Twitter[1] users and the users' perspectives on the relationship between science-based reasoning and justified belief. Two perspectives are distinguished: one is the "pro-science" perspective, and the other is the "non-pro-science" perspective. These are defined in terms of science-based[2] and non-science-based beliefs, which are in turned defined in terms of science-based reasoning. Science-based reasoning is explained in Section 1.1.

1.1 The Meaning of "Science-Based"

1.1.1 The Meaning of "Science": The Demarcation Problem

Defining "science-based" and "non-science-based" precisely is more difficult than it seems at first blush for several reasons, one being that science is hard to define. The task of demarcating science and non-science (which is part of the larger task in philosophy to determine which beliefs are epistemically warranted (Hansson, 2014)), is often referred to as "the demarcation problem". This term if often meant to include the problem of further demarcating different kinds of non-science

[1] Twitter is a micro-blogging website where users (i.e., people who have accounts on the site) post tweets (also called *status updates*), which are essentially short blog posts, limited to 140 characters.

[2] My usage of the term "science-based" as distinct from "evidence-based" is borrowed from the blog *Science-Based Medicine*'s "About" page by Novella (2013).

(e.g. Hansson (2014) mentions “unscience” and pseudoscience[3]). It is a notoriously difficult task, and attempts to provide definitions for the various categories have been been controversial (Pigliucci and Boudry, 2013; Hansson, 2014). Fortunately, despite the difficulty of this abstract definitional problem, and disagreement among commentators regarding how it is to be solved, there is remarkable agreement among philosophers and scientists when it comes to identifying specific examples as science or pseudoscience (Pigliucci and Boudry, 2013; Hansson, 2014). The definition of science-based depends on the current state of science, and not necessarily on the definition of science, which precludes the demarcation problem from undermining the validity of the concept of science-based reasoning. The following sections discuss evidence-based reasoning in general (i.e. forming beliefs from evidence), and then science-based reasoning, which is a specific kind of evidence-based reasoning.

1.1.2 Evidence-Based Reasoning

The idea that beliefs should be evidence-based is an old one. Hume (2012) says “a wise man . . . proportions his belief to the evidence” (p. 122). A slightly expanded and more explicit formulation comes from Quine and Ullian (1978):

> Insofar as we are rational in our beliefs, . . . the intensity of belief will tend to correspond to the firmness of the available evidence. Insofar as we are rational, we will drop a belief when we have tried in vain to find evidence for it. (p. 16)

This conception of the relationship between evidence and rational belief is sometimes called “Evidentialism”. Evidence-based reasoning involves tacit agreement with Evidentialism.

Blanshard (1974) touches more explicitly on an important point about lack of belief: “[the scientific] ideal is to believe no more, but also no less, than what the evidence warrants” (p. 411). One’s belief system or perspective on reality is not necessarily evidence-based just because all of one’s beliefs are motivated by and proportional in intensity to one’s evidence. That is necessary but not sufficient. In order to have an evidence-based belief system, one *must* form beliefs in the face of evidence. A lack of belief can be, in that sense, non-evidence-based. So, while

[3] I use the term “pseudoscience” to refer to non-science that is presented as if it were science.

the term "beliefs" is used throughout this thesis, in fact what is meant is "belief *statuses*," i.e., either a belief or lack of belief.

Crucially what matters in evidence-based reasoning is not any particular bit of evidence E, but rather *total evidence* (Kelly, 2008). E, considered by itself, might justify believing that hypothesis H is true, but it may be that there is some other evidence E', such that, taken together, E and E' do not justify believing H (Kelly, 2008). For example:

> Even if I am initially justified in believing that your name is Fritz on the basis of your testimony to that effect, the subsequent acquisition of evidence which suggests that you are a pathological liar tends to render this same belief unjustified. (Kelly, 2008)

Kelly goes on to distinguish two ways in which some evidence E' can "defeat" the status of E as evidence bearing on H: undercutting and rebuttal. In the example about deciding whether or not to believe the man's name is Fritz, we saw undercutting; the second bit of evidence suggested that the first bit was not a reliable indicator of the man's name; that is, it undermined the idea that there is an evidential connection between E and H at all). In contrast, rebuttal is when some E' supports the falsity of H (i.e. supports $\neg H$) more strongly than E supports H (Kelly, 2008). In this example, perhaps seeing that both his driver's license and passport display the name "John" would successfully rebut the evidence from the man's own testimony. The idea of defeating evidence, and the idea of a belief's intensity being determined by the strength of evidence, both rely on the idea of some evidential connections being stronger than others. I will not go into detail on this topic. However, it is important, since in this thesis I construe a belief as evidence-based only if its intensity is commensurate with the strength of the evidence, as determined by the strength of the evidential connection between the evidence and the content of the belief (e.g. a firm belief arising from incommensurately weak evidence is not truly evidence-based, even if the belief, construed in binary form, is more evidence-based than the denial of the content of the belief). Regarding why certain observations or artifacts are not considered "good" evidence for something paranormal or supernatural, it is often helpful to keep in mind the phenomenon in which evidence is defeated. Undercutting is particularly important for understanding when it comes to beliefs that amount to unfalsifiable claims. For example, it is hard to conceive of any observation or data that could rebut someone's testimony that there is some incorporeal being that may reveal itself to people in an unpredictable fashion.[4] However, it can be undercut by the observation that people sometimes say things that are not true. Therefore, the testimony does not necessarily make belief in the described being evidence-based (but, arguably, it might make the belief evidence-based, depending on one's other evidence[5]).

Something can only be evidence for something else to the extent that there is an evidential connection between the two (Kelly, 2008). Evidential connections are determined by one's "background theory" about the way the world works (Kelly, 2008). For instance, according to current scientific theory, acidic liquids cause litmus paper to turn red; therefore, according to current scientific theory, litmus paper turning red in a liquid is extremely strong evidence that the liquid is acidic. If one were to assume an alternative theory in which color change in litmus paper is determined only by the presence of sodium ions in a liquid and not pH, then there would be no evidential connection between the color that the litmus paper turns when dipped in a liquid and the acidity of the liquid. In principle, evidence-based reasoning can be used with any background theory.

1.1.3 Science-Based Reasoning: Jointly Reasoning from Science and Evidence

Science-based reasoning is a form of evidence-based reasoning in which one's background theory, which determines the strength of evidential connections, is current scientific theory, and in which hypotheses are always considered in the context of the most plausible alternate hypotheses.[6] An example of a science-based belief would be the firm belief that microorganisms exist, or that iron and aluminum are made up of different elements. In this case, the denial of the content of the belief would

[4] That is to say, it is hard to imagine anything that would constitute evidence for the hypothesis that such a being does not exist.

[5] This fact is why evidence-based reasoning alone is not necessarily sufficient for separating the previously mentioned non-science-based beliefs from science-based ones. Appeal must be made to scientific plausiblity in order to make that separation.

[6] Science-based reasoning may involve rejection or alteration of aspects of current scientific theory, given proper consideration of the level of support (in terms of evidence) for the aspects of the theory to be altered or rejected vs. the support for the change or rejection. This is the process by which scientific theory evolves.

be a non-science-based belief status. A more every-day example might be someone's belief that their friend is angry with them. So long as the believer followed the reasoning process just described, the belief is science-based. In this case, denial of the content of the belief would probably also be science-based, depending on the denier's evidence, since it is generally scientifically plausible for someone to be either angry or not angry at any other person. In this thesis, I am generally less interested in beliefs about such mundane or everyday questions, and more interested in claims and beliefs that are controversial (examples are in Section 1.2). Note that, in many cases, what might be considered an everyday, mundane non-science-based belief (e.g., that a god is responsible for one's dinner, or that karma is the reason someone lost their wallet), in fact entails, presupposes, or implies a more general highly controversial belief, (e.g. belief in a god, or belief in the reality of karma).

Science Requires Science-Based Reasoning

In some cases, the flaw in would-be scientific studies that causes them to be categorized as pseudoscience is just the failure of researchers to ensure that their hypotheses and interpretations of results exclusively involve science-based reasoning rather than some other kind of (possibly evidence-based) reasoning. Examples from recent history include ganzfeld[7] studies (e.g., those discussed by Hyman and Honorton (1986)), and, still more recently, Bem (2011), which all fall under the umbrella of parapsychological (or "psi") research. These studies all purported to find evidence that suggests that psi abilities are real. If we had no reason, based on the methodology outlined by the researchers, to suspect that the evidence resulting from the studies in question was gotten using a scientifically inappropriate or otherwise problematic methodology,[8] the experimental results, according to science-based reasoning, still do not justify the belief that experimental subjects used psi abilities. The reason for this is that, while the reality of psi abilities would explain the results, the proposition that psi abilities are real is so scientifically implausible that, according

[7] Ganzfeld experiments involve sensory deprivation. The idea in psi research is that extra-sensory perceptions might be easier to focus on in such a state, since sensory input could drown out extra-sensory perceptions.

[8] This is highly controversial. Hyman and Honorton (1986) say that there were methodological problems in the ganzfeld studies done up to that point, and say that, as a result, the those studies are inconclusive regarding the reality of psi abilities. Doubt is cast on Bem (2011)'s results by Galak *et al.* (2012)'s failure to replicate Bem (2011)'s results.

to science-based reasoning, the best hypothesis would be that there are yet-unknown methodological problems in the studies, and potentially that random chance is at least partially responsible. However, the researchers from the previously mentioned ganzfeld studies and Bem (2011) did say that their respective results suggest that psi abilities are real, and such failures to use science-based reasoning preclude us from categorizing their research as real science.

1.2 Examples of Prototypical Non-Science-Based Beleifs

Prototypical examples of non-science-based beliefs that are focused on in this thesis include the following:

- belief in the existence of
  - ghosts (Blackmore and Moore, 2014)
  - extra-sensory perception (Goode, 2013)
  - supernatural beings (Fishman, 2009)
  - Bigfoot and other cryptids (Loxton and Prothero, 2012)
- subscription to Creationism (Hansson, 2014)
- belief in the efficacy of
  - homeopathic remedies (Shang *et al.*, 2005; Hansson, 2014)
  - prayer (Benson *et al.*, 2006)
- belief in the predictive power of astrology (Hansson, 2014)
- belief that global warming is
  - not happening (National Research Council Committee on America's Climate Choices, 2011)
  - unrelated to anthropogenic greenhouse gases (National Research Council Committee on America's Climate Choices, 2011)
- belief that vaccines cause autism (DeStefano *et al.*, 2013)

1.3 Perspectives on Science-Based Reasoning and Rational Belief

I use science- and non-science-based belief to define epistemological *perspectives*. From the “pro-science” (PS) perspective, only science-based beliefs are rational, but from the “non-pro-science” (NS) perspective, other kinds of beliefs are also rational. The terms PS and NS may also be used adjectivally to refer to people with a PS or NS perspective. An important assumption in this thesis is that people only express (via assertion, implicature, or presupposition) beliefs that they consider rational. That is, anyone with the PS perspective only expresses science-based beliefs, and expression of a single non-science-based belief is sufficient to preclude someone from the pro-science category. Since people are not compelled to express all of their beliefs, it may be that someone who expresses only science-based beliefs is in fact NS, but has, for whatever reason, not expressed any of their non-science-based beliefs. This is a problem for any attempt to conclusively identify someone as PS. The way this issue is dealt with in this thesis is discussed in Section 2.1.3.

1.4 Theses Proposed

Thesis 1. *The simple topic alignment technique described in Section 3.1.2 produces plausible representations of differences in how PS and NS Twitter users write about the same underlying real topics.*[9]

In Section 3.1.2 I propose a straightforward technique of comparing the language of topically similar datasets. It involves creating topic models for the two datasets and then uniquely aligning the topics such that the overall divergences between matched topics are minimized (intuitively, this just means the best unique matching of topics). In Section 4.2, I give the matched topics and discuss their apparent quality (according to impressions from examining of the topics' top words), and discuss what they suggest about the PS and NS data.

Thesis 2. *The linguistic differences between PS and NS user documents are robust enough that accurate automated classification is possible using only linguistic*

[9] Here, “real” topics are topics in the normal sense, rather than the LDA sense; see Section 3.1.2 for more information.

features.

In Section 4.1 word-frequency measure comparisons between the PS and NS datasets are given, which provide linguistically quantifiable differences between the PS and NS documents. But a clearer empirical validation for the claim of reliable linguistic difference is good classification results using textual linguistic features. In Chapter 5, documents from the PS and NS datasets are classified with high accuracy using such features. To support the idea that PS and NS perspectives cut across topics, classifiers are trained and tested on different topical subsets of the data (see Section 5.4). By providing a comparision of the results from basic classification to results from classification in which documents are stripped of most retweets (see Section 5.2), high classification accuracy is shown not to depend on the presence of retweets.[10] Other experimental results support the idea that higher frequency of a user's perspective-revealing tweets (see Section 5.6), and higher extremeness of beliefs (see Section 5.5), makes them more likely to be classified correctly.

A weakness of supervised classification is the need for labeled documents, which, for many tasks, forces researchers to employ human annotators to label documents manually. Researchers like Thamrongrattanarit *et al.* (2013) have attempted to make labeling inexpensive and quick by taking advantage of easy ways to get potentially noisy, but decent-quality labels, without resorting to individually evaluating the documents. User-curated lists of other users on Twitter, known as "Lists,"[11] are used to similarly label large numbers of users easily (explained more fully in Section 2.2). Using classification results from training on noisy labels, the results of which are in Section 5.7, distant supervision is shown to give accuracies worse than full supervision, though still potentially useful, depending on the goal.

1.5 A Potential Application

Pro-science writers (e.g. Sagan (1996)) and bloggers (e.g. Dunning (2012)) have expressed interest in systems for identifying non-science-based claims. They em-

[10] A "retweet" (also written "RT") is a tweet that is simply a quotation of an entire tweet written by someone else, with or without commentary. "Automatic retweets" contain no added commentary, but "manual retweets" generally do contain commentary.

[11] On Twitter, "Lists" are simply lists created by users to which other users are added as members. These Lists often have a theme or topic (or perspective).

phasize that pseudoscientific ideas should be avoided not only because they are unsupported by scientific evidence, but because they can lead to wasted time and resources, and they even pose dangers to the public (Hansson, 2014; Pigliucci and Boudry, 2013). These commentators generally also express belief that non-science-based beliefs are inherently undesirable or unreasonable. This thesis does not take a position one way or the other about the inherent value of science-based reasoning and non-science-based reasoning, only the position that science-based reasoning should be used when considering science-related issues. But in any case, there is interest from these commentators in detection of science-based and non-science-based claims generally.

This thesis could provide utility even for the general population in connection with those kinds of non-science-based beliefs that are demonstrably dangerous. Unfounded claims or beliefs regarding health can be dangerous. The website `whatstheharm.net` documents cases where ideas from pseudoscience or other kinds of non-science have had a negative effect someone's life (ranging from examples of people wasting money on ineffective treatments to instances of injury or death).[12] Cases are cited where people in need of medication or other medical treatment instead turned to "alternative medicine", and died.[13] The classifier made in this thesis might help identify Twitter users who are likely to have science-based ideas about these topics. Another dangerous non-science-based idea is climate change denial (Hansson, 2014; Pigliucci and Boudry, 2013), which could harm members of future generations. For this reason, identifying people likely to propagate climate change denial could be useful when attempting to form a science-based opinion regarding appropriate policy. Misunderstandings about science are also important in court cases, when people evaluate evidence and interpretations of evidence presented by lawyers and experts (Pigliucci and Boudry, 2013). Regarding science-related issues or questions in general, identifying people who are dedicated to science-based reasoning is useful as an indication of credibility.

[12] The site seems well done, but I cannot vouch for its total accuracy, and it seems generally biased toward science-based thinking.

[13] The page dedicated to alternative medicine is `whatstheharm.net/alternativemedicine.html`

Chapter 2

Data

In order to study any linguistic differences between text generated by PS and NS writers, four datasets were compiled, three being made up of Twitter data, and one being made up of blog posts and articles (henceforth I use "article" to mean article[1] or blog post).

2.1 The Main Dataset

The "main" dataset is a collection of documents, where each document has a label and the most recent tweets of some user. The first step towards creating this dataset

[1] Note that"article" not in the peer-reviewed journal sense, but in the newspaper or magazine sense.

label(s)	PS	NS	total
health	372 (28.2%)	112 (8.5%)	484 (36.7%)
health + relig	16 (1.2%)	21(1.6%)	37 (2.8%)
health + paranormal	5 (0.4%)	3 (0.2%)	8 (0.6%)
health + relig + paranormal	0 (0.0%)	5 (0.4%)	5 (0.4%)
relig	223 (16.9%)	334 (25.3%)	557 (42.3%)
relig + paranormal	8 (0.6%)	39 (29.6%)	47 (3.6%)
paranormal	26 (2.0%)	154 (11.7%)	180 (13.7%)
total	650 (49.3%)	668 (50.7%)	1318 (100.0%)

Table 2.1: Distribution of users over label combinations in the main dataset

	PS	NS	total
total documents	650	668	1318
total tweets	714,461	714,715	1,429,176
total non-RTs	548,522	583,187	1,131,709
mean tweets/document	1,099.2	1,069.9	1,084.4
mean non-RTs/document	843.9	873.0	858.7
proportion of non-RTs	0.77	0.82	0.79
smallest number of tweets in a doc	107	102	102
largest number of tweets in a doc	1,199	1,199	1,199

Table 2.2: Basic information about the main dataset

label(s)	PS	NS	total
health	261 (28.5%)	79 (8.6%)	340 (37.1%)
health + relig	9 (1.0%)	14 (1.5%)	23 (2.5%)
health + relig + paranormal	0 (0.0%)	3 (0.3%)	3 (0.3%)
relig	158 (17.2%)	234 (25.5%)	392 (42.7%)
relig + paranormal	4 (0.4%)	25 (2.7%)	29 (3.2%)
paranormal	18 (2.0%)	112 (12.2%)	130 (14.2%)
total	450 (49.1%)	467 (50.9%)	917 (100.0%)

Table 2.3: Distribution of users over label combinations in the training set (subset of the main dataset)

	PS	NS	total
total documents	454	468	922
total words	6,623,447	6,790,592	13,414,039
total tweets	505,067	498,212	1,003,279
total non-RTs	386,850	408,132	794,982
mean tweets/document	1,112.5	1,064.6	1,088.2
mean non-RTs/document	852.1	872.1	862.2
proportion of non-RTs	0.77	0.82	0.79
smallest number of tweets in a doc	178	102	102
largest number of tweets in a doc	1,199	1,199	1,199

Table 2.4: Basic information about the training set (subset of the main dataset)

label(s)	PS	NS	total
health	56 (28.4%)	17 (8.6%)	73 (37.1%)
health + relig	2 (1.0%)	3 (1.5%)	5 (2.5%)
health + relig + paranormal	0 (0.0%)	2 (1.0%)	2 (1.0%)
relig	34 (17.3%)	50 (25.4%)	84 (42.6%)
relig + paranormal	3 (1.5%)	6 (3.0%)	9 (4.6%)
paranormal	3 (1.5%)	21 (10.7%)	24 (12.2%)
total	98 (49.7%)	99 (50.3%)	197 (100.0%)

Table 2.5: Distribution of users over label combinations in the dev set (subset of the main dataset)

	PS	NS	total
total documents	98	100	198
total words	1,392,025	1,516,432	2,908,457
total tweets	104,977	109,302	214,279
total non-RTs	79,452	88,346	167,798
mean tweets/document	1,071.2	1,093.0	1,082.2
mean non-RTs/document	810.7	883.5	847.5
proportion of non-RTs	0.76	0.81	0.78
smallest number of tweets in a doc	107	131	107
largest number of tweets in a doc	1,199	1,199	1,199

Table 2.6: Basic information about the dev set (subset of the main dataset)

label(s)	PS	NS	total
health	55 (28.1%)	16 (8.2%)	71 (36.2%)
health + relig	5 (2.6%)	4 (2.0%)	9 (4.6%)
health + relig + paranormal	0 (0.0%)	0 (0.0%)	0 (0.0%)
relig	31 (15.8%)	50 (25.5%)	81 (41.3%)
relig + paranormal	1 (0.5%)	8 (4.1%)	9 (4.6%)
paranormal	5 (2.6%)	21 (10.7%)	26 (13.3%)
total	97 (49.5%)	99 (50.5%)	196 (100.0%)

Table 2.7: Distribution of users over label combinations in the test set (subset of the main dataset)

	PS	NS	total
total documents	98	100	198
total words	1,387,578	1,379,618	2,767,196
total tweets	104,417	107,201	211,618
total non-RTs	82,220	86,709	168,929
mean tweets/document	1,065.5	1,072.0	1,068.8
mean non-RTs/document	839.1	867.1	853.2
proportion of non-RTs	0.79	0.81	0.80
smallest number of tweets in a doc	133	260	133
largest number of tweets in a doc	1,199	1,199	1,199

Table 2.8: Basic information about the test set (subset of the main dataset)

	docs	words
PS	99	142,098
NS	101	92,781
total	200	234,879

Table 2.9: Number of documents in the article dataset with each perspective.

involved finding and labeling users. The labels given to users in fact contain sub-labels, namely, one perspective label and one or more topic labels. The information encoded in the labels consists entirely of subjective judgments. Sections 2.1.1 and 2.1.1 and 2.1.2 describe the basic semantics of all the parts of the labels, and Section 2.1.3 gives the guidelines used for creating the perspective and topic labels for users; understanding the guidelines that were used is needed to fully understand the meaning of the labels, which is particularly important regarding the categorization of users as PS vs. NS.

2.1.1 Perspective Label

This label has the following format: "perspective-extremity-frequency."[2] There are two possible perspectives: PS (pro-science) and NS (non-pro-science). The "extremity" part of the label encodes "extremity of perspective" and the "frequency" part

[2] The first users that were labeled lacked extremity and frequency information in their labels: the only useful information they contain is a perspective label. The result was that 318 user documents in the main dataset lack extremity and frequency labels.

queries
“believe in ghosts”
“don’t vaccinate”
#atheist OR #atheism
#autism AND #vaccine
#wakefield #fraud
(gingko OR ginkgo) AND focus
UFO AND sightings
astrology
bigfoot AND myth
creation AND science
god AND science
herbal AND #BigPharma
quackery
religion AND #Science
soul AND energy
soul AND scientific AND evidence
spiritual AND energy

Table 2.10: Examples of queries that were used to find Twitter users. These give a sense of the kinds of topics that are represented most prominently in the data, such as alternative medicine, gods, and spirituality.

of the label encodes a rough estimate of frequency of perspective-revealing tweets.

The extremity label is meant to give some sense of how enthusiastic users appear to be about their perspective. This is meant to be an independent measure from frequency of perspective-revealing tweets. In fact, extremity labels were often based on a single tweet (a user need not tweet often on an issue to demonstrate high extremity of perspective on the issue) (see Section 2.1.3 for information on how these labels were chosen). The extremity labels are a bit unintuitive: the meaning is slightly different depending on whether they are a label for PS or NS. For PS, the extremity label refers to their enthusiasm for science-based reasoning, whereas for NS, the extremity label indicates enthusiasm for specific types of non-science-based reasoning (e.g., reasoning about scientific questions such as earth's formation based on the Bible), or even just particular non-science-based claims (e.g., that vaccines cause autism). The use of perspective-dependent semantics of the extremity label is a result of the observation that it is not likely that anyone is an enthusiast of non-science-based reasoning in general (which would include every other kind of evidence based reasoning, along with non-evidence based reasoning). Extremity labels are on a scale from 1 to 5, with increments of 0.5.[3]

The frequency (which is a rough estimate of the proportion of tweets that reveal the user's perspective) is on a scale from 0 to 6, with increments of 0.5. The label 0 indicates that the user's perspective does not come across in any tweet; that label was only used when the user's "bio" (i.e. self description on their home page) indicated a perspective, but no tweets were found that did.[4] This subjective frequency label was used because it was much faster to determine these subjective labels than it would have been to use an objective measure. As discussed in Section 2.6, the frequency and extremity are not used in such a way that spending additional time employing a more rigorous system was warranted.

For both extremity and frequency labels, whole numbers were almost always used, even though increments of 0.5 were possible. This was simply due to the fact

[3] The meanings of the whole number points on the scale are as follows: 1 means "barely on the side of the indicated perspective"; 2 is low extremity, 3 is medium (neither low nor high) extremity, 4 is high extremity, and 5 is an advocate of their perspective.

[4] The meanings of the whole number points on the scale are as follows: 0 means "bio only" (as previously stated), 1 means there was one tweet on the side of the indicated perspective; 2 means that the user rarely revealed perspective. 3 is low frequency (of perspective-revealing), 4 is medium (neither low nor high) frequency, 4 is high frequency, and 5 is extremely high frequency.

that it seemed that increments of 0.5 were unnecessarily fine-grained.

2.1.2 Topic Labels

There could be any number of topic labels for a given user. Topic labels have the following format: "topic-frequency-[subtopic-frequency],"[5] where the "[subtopic-frequency]" portion could be a list of any number of "subtopic-frequency" elements. Ultimately, subtopics were scarcely made use of in this thesis. The topic frequency label roughly indicates some sense of the frequency with which the user tweeted about the topic in question and is on a scale from 0 to 6. These labels are closely analogous to the frequency labels within the perspective label discussed in Section 2.1.1.[6] Subtopics also had their own frequency labels, which were exactly analogous to the topic frequency labels. There could be any number of topic labels for a given user.

During the labeling process, a large number of topics and subtopics were used, and new ones were added whenever new topics/subtopics of interest were encountered. However, topic labels, and the subtopic labels they contain, are not meant to be exhaustive. That is, the absence of a topic label does not necessarily mean the user never tweets on the topic.

By far the most common topic labels are "health" and "religion," which is evident in Table 2.7. Minor topics include "climate change" and "psychic". After data was collected, all minor labels were either ignored for all experiments (e.g. the "climate change" label), or grouped under the *post hoc* label "paranormal." No user was labeled such that their only topic label was a minor label that was not ultimately subsumed by "paranormal." The end result was that everyone had at least one of the following labels: "religion" (any kind of spirituality), "health" (e.g., alternative medicine, vaccines, praying for health), or "paranormal" (e.g., astrology, psychic abilities, ghosts, extraterrestrials, and cryptozoology).

[5] The first users that were labeled lacked frequency information in their topic and subtopic labels: the only useful information they contain is the identity of the topic. The result was that 318 user documents in the main dataset lack topical frequency labels.

[6] In this case meanings of the whole number points on the scale are as follows: 0 means "bio only" (as previously stated), 1 means there was one tweet on the indicated topic; 2 means that the user rarely tweeted on this topic. 3 is low frequency of tweets on the topic, 4 is medium (neither low nor high) frequency, 4 is high frequency, and 5 is extremely high frequency.

2.1.3 The Guidelines for Creating Labels for Users

This section describes how the definitions of PS and NS given in Section 1.3 were operationalized in the creation of the annotations for a given user. The process of finding users to annotate is described in Section 2.1.5.

Tweets of users to be labeled were skimmed for approximately three minutes, starting from the most recent. In no case were all of the user's tweets read thoroughly (most users had over 1,000 tweets). If any tweets were seen that revealed any non-science-based belief of any kind, the user would be considered NS in connection with those topics. The extremity of belief label for NS (i.e., within the perspective label) would be based on the most extreme NS tweet on any topic.

The frequency label would be based on the estimated frequency of tweets that revealed non-science-based beliefs on any topic. Seemingly science-based beliefs about mundane, uncontroversial things (e.g. a belief of the user that he or she has just eaten a sandwich) had no effect on the label the user received.

If in any tweets, any science-based beliefs are expressed on topics of interest (e.g. anything supernatural or paranormal; anything in religion; alternative medicine, vaccination, including vaccinations, homeopathy, some herbal supplements and vitamin supplements; generally anything that is controversial in popular culture), this user was considered PS in connection with those topics.

If it was found that the user is identified as PS and NS in connection with either the same topic or with different topics, the user was considered conflicted, and never used in any part of this thesis. This is discussed in Section 2.6.

Examples of Tweets and the Labels they Received

Here is an example of a tweet that resulted in a PS label with an extremity of 2 ("low extremity"): "I don't believe in the whole astrology/horoscope thing. Seems a little far fetched, I don't know." Contrast with the following example of a tweet resulting in a PS extremity label of 5 ("advocate"): "there is no debate about creationism. evolution is a fact. creation belongs in myth, theology and folklore curricula, not science education." The following is an example of a low-extremity NS tweet (extremity of 2): "After every heartbreak, the art of music heals and restores my soul." Contrast with the following example of a tweet resulting in an NS extremity label of 5 ("advocate"): "We must not only evolve our SPIRITUALITY, however we must learn & overstand the

Matrix which we ALL live in day to day."

Only 16 "conflicted" users were found. One user expressed a science-based view of vaccines (namely, the view that people should get them) (extremity level of 3, or "medium extremity") in the following tweet: "Just did them too RT @Wh1t3Rabbit: TDAP & Flu vaccines done. Need it for the babies, so had to suck it up and get poked. #WorthIt #DadOps." However, they also expressed a non-science-based belief regarding religion/God (extremity level of 4, or "high extremity"): "@branthansen the other thing regarding Austin's story is that if he quits witnessing, he allows the sin to be a wedge between him and God." This user, like all conflicted users, was excluded from the data. Note however that enthusiasm for science expressed in connection with topics that were not of interest (e.g., photosynthesis in plants) would not be sufficient for labeling someone PS, and therefore does not amount to a conflict when paired with non-science-based beliefs on other topics.

2.1.4 Criteria in Deciding Whether to Label a User

Not every user whose profile was examined received a label. Some kinds of users were considered undesirable. Anyone with fewer than 100 tweets was excluded, as was anyone who had not tweeted in the past month. The preferred accounts for labeling were ones where most tweets are manually generated, made up of normal language, and not advertisements or spam. Here is an example of a tweet that is an advertisement, or spam-like: "#GingkoBiloba& #BilberryLeaf are in our #BoggleBooster!You can add it #Free to reg sizeClassic,#Superfood orSpirit smoothie #yummyandhealthy". And here is an example of a tweet that is not normal language, and clearly not manually generated: "Photo: yellowmeece: therealkillthetraitor: deadbilly: therealkillthetraitor: xion1212." Neither of the users that posted these received a label. The problem with such automated tweeting is that automatically generated content usually has a stock phrasing or uses the same word every time. Automatically generated content, exemplified in the latter tweet, was acceptable in small amounts (the decision was based on subjective impression of the frequency automatic tweets in the user timeline).[7] The latter example tweet seems to have been generated by software that always starts tweets out with the word "photo;" if half of someone's tweets are

[7] A user timeline is the list of tweets (including retweets) that have ever been posted by a user. It is shown on a Twitter user's profile page, where the list of tweets is in reverse chronological order.

automatically generated in this way and therefore start with “photo”, then this is going to have an explosive affect on the frequency of that word compared to what the frequency would have been if the automatic tweets were not there. This is not desirable, because less can be learned about real, human-generated language if the dataset contains users whose timelines are full of repetitive automatic tweets. A final reason not to a label a user is if no perspective has been identified after three minutes.

2.1.5 The Process of Finding Users to Label

The process of finding and labeling users involved querying Twitter using queries specifically designed to find tweets that reveal PS or NS perspectives on topics of interest. A few example queries are given in Table 2.10, and the full list of queries is in Appendix A. Some queries were targeted to find users of a specific perspective (e.g., *quackery* was aimed at PS users), while others were not (e.g., *god AND science*). The process was as follows:

1. Search Twitter using a query that will probably return tweets revealing a PS or NS position on a topic of interest.

2. If you see a tweet that seems PS or NS and that was posted by a user who has not already received a label, go to the timeline of the user who posted it. Otherwise go back to Step 1.

3. Spend approximately three minutes skimming tweets. If the user appears to be PS or NS, create a label for the user and save the label and screen name in a document. Go back to Step 2.

In step 2, there is the option not to label the user. Reasons for not labeling a user are discussed in Section 2.1.4.

Occasionally, once I found an interesting user, I looked at the Lists that they were a member of in hopes of finding more good candidates. The process of finding users this way is described here:

1. Go to the account page of a user who has been identified as PS or NS and whose List memberships have not already been examined. If there is a yet-unseen List whose title or description suggests it may be relevant, go to its timeline. Otherwise repeat this step.

2. If you see a tweet in the List timeline that seems PS or NS and that was posted by a user who has not already received a label, go to the timeline of the user who posted it. Otherwise go back to Step 1.

3. Spend approximately three minutes skimming tweets. If the user appears to be PS or NS, create a label for the user and save the label and screen name in a document. Go back to Step 2.

The vast majority of users that were individually labeled were found using queries, rather than using List timelines.

2.1.6 Creation of a Document for Each Labeled User

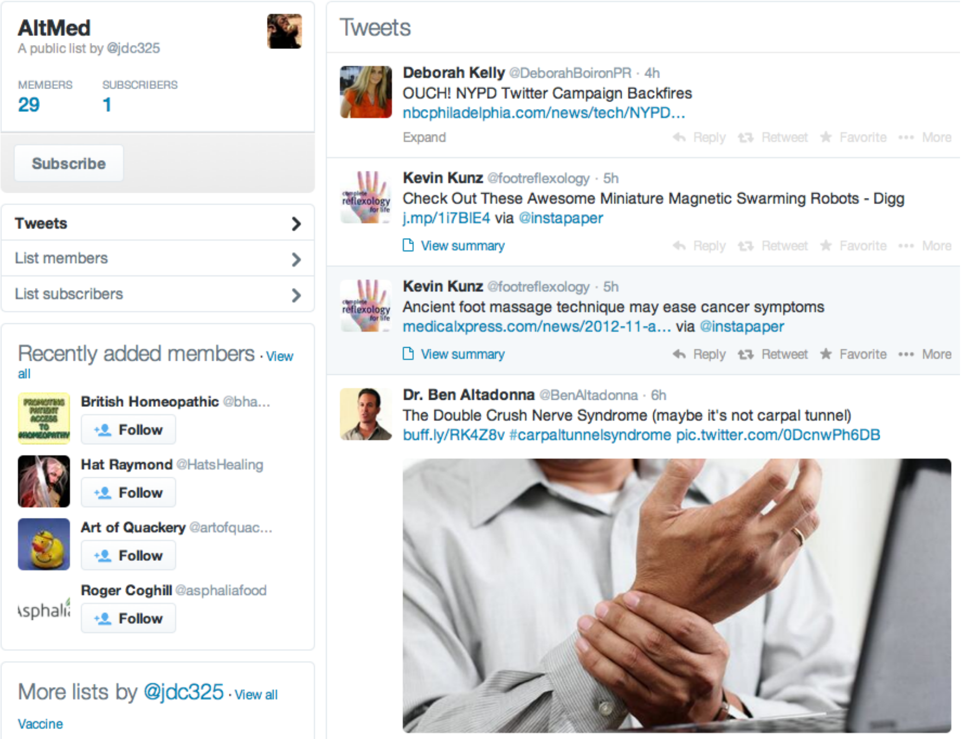
After what was judged to be sufficiently large list of labeled users had been compiled, the next phase of creating the main dataset began. For each user in the set of labeled users, a document was created that contained their most recent tweets. The exact number of tweets collected for a user ranges from 102 to 1,199.[8] The average number of tweets is 1084.4. More data summarization is in Table 2.9 Whenever a user document contains fewer than 1,000 tweets, it is because the user had not yet tweeted 1,000 times before the user document was created. Approximately 1,400 users received labels, but due to accounts being deleted and accounts becoming private after being labeled but before the document creation stage, only 1,318 user documents could be created.

2.2 The Noisy Dataset: Weakly Labeling Users via Lists on Twitter

Another set of users was collected and labeled by individually labeling Lists. The way people interact with Lists is by viewing the List's timeline.[9] The reason users

[8] Originally I intended to collect anywhere between 100 and 1,000, but usually, Twitter gave me "extra" tweets (up to 199 extra), and I chose to keep them rather than to discard them. This was because I continued requesting pages of 200 tweets until I reached 1000, but often pages contained a few less than 200. So there were cases where I ended up requesting 200 tweets even though I was at 999. And so in that case I included the excess 199 tweets in the document.

[9] For a List, a timeline is where all the tweets of all List members appear, ordered by decreasing recency.



AltMed
A public list by @jdc325

MEMBERS 29
SUBSCRIBERS 1

Recently added members · View all

British Homeopathic @bha...

Hat Raymond @HatsHealing

Art of Quackery @artofquac...

Roger Coghill @asphaliafood

More lists by @jdc325 · View all

Vaccine

Tweets

Deborah Kelly @DeborahBoironPR · 4h
OUCH! NYPD Twitter Campaign Backfires
nbcphiladelphia.com/news/tech/NYPD...

Kevin Kunz @footreflexology · 5h
Check Out These Awesome Miniature Magnetic Swarming Robots - Digg
j.mp/1i7BIE4 via @instapaper

Kevin Kunz @footreflexology · 5h
Ancient foot massage technique may ease cancer symptoms
medicalxpress.com/news/2012-11-a... via @instapaper

Dr. Ben Altadonna @BenAltadonna · 6h
The Double Crush Nerve Syndrome (maybe it's not carpal tunnel)
buff.ly/RK4Z8v #carpaltunnelsyndrome pic.twitter.com/0DcnwPh6DB

Figure 2.1: Example of a view of a List. This one recieved an NS perspective label and a "health" topic label, labels which was subsequently inherited by all of its members.

make Lists is that Lists make it easy to read all the tweets of a specified group of users. They often have a theme to them (e.g. one can find Lists of users who tweet a lot about sports). An example of a List timeline is in Figure 2.1. Twitter Lists were used to easily acquire lists of users that are likely to have a certain perspective on a certain topic. The idea is that a List could be labeled, and then the List label could be applied to everyone in the List. The labels given to users in this way are likely to be noisy.

The guidelines for labeling Lists were identical to those for labeling users (as outlined in Section 2.1.3), where the List is treated as if it were a user (i.e. as if the List were a user with an identifiable perspective). The process used to find and label Lists is summarized as follows:

1. Go to the home page of a user that received an individualized label as described in Section 2.1, and whose Lists have not already been searched for Lists that might be good candidates for labeling.

2. Go to the Lists that the user is a member of. If there is a yet-unseen List whose title or description suggests it may be relevant, go to its timeline. Otherwise go back to Step 1.

3. Spend approximately three minutes skimming the tweets in the List's timeline. If it seems that this List is a good candidate, create a label for the List and save the label, List name, and screenname of the List creator in a document.[10] Go back to Step 2.

2.2.1 Creation of Documents for Weakly Labeled Users from Each Labeled List

After 25 Lists had been labeled, the next step toward creating a set of weakly labeled user documents involved associating the List labels with the screen names of the List members. Rather than do this for all members of the List, only the first 200 members were used (unless the List had fewer than 200 members, in which case all users were used); the purpose of limiting the number to 200 was to stop huge Lists from drowning out smaller Lists in the final dataset. So for each labeled List, a set of weakly labeled users was produced (users who were in multiple Lists kept whatever weak label they received first in this process), yielding one large set of weakly labeled users. The final step, wherein documents of the most recent tweets were created for each user, was the same as what is described in Section 2.1.6, except instead of using the list of individually labeled users, the list of weakly labeled users was used.

2.3 Streaming Random Tweets

A document containing over a million tweets was made by streaming a random sample of tweets from Twitter and discarding any that were not labeled by Twitter as being in English.

[10] The List name and creator screen name are both needed to access a List or List timeline with the Twitter API.

2.4 The Article Dataset

A set of articles was collected and labeled. The semantics of the labels is extremely similar to that of Twitter user labels (described in Section 2.1). However, due to the inherent difference between a Twitter user and an article, there is a difference in the meaning of the "frequency" part of the label. In the case of articles, the frequency label is a rough estimate of the proportion of sentences, rather than tweets, that reveal the perspective in question or touch on the topic in question. Additionally, I only labeled articles based on the text content of the article proper, not any metadata connected to the article or other information I had about the author. As a result, a frequency label of zero was impossible for articles: if a perspective were revealed in no sentences, the article would not have been labeled at all. The websites that the articles came from, and the number that came from each, is given in Appendix B. One author, identified only as "Alise" in her blog, has 16 articles in the dataset. Everyone other author represented has fewer than 15.

2.5 How the Various Datasets are Used

The main dataset is used for the fully supervised Twitter user document classification tasks in Chapter 5, i.e., all classification tasks other than those in Sections 5.7, in which the noisy dataset is used for training (though the main dataset is still used for evaluation), and 5.8, in which the main dataset is used for training, but the classifier is evaluated on the article dataset (the set of articles is small, and it is only used for testing). The main dataset is also the one that is analyzed in Chapter 4, which involves calculating word frequencies and making topic models.

The document made from the random stream of Tweets was used as a neutral dataset for calculating frequency ratios. For every word that appeared in the PS, NS, or neutral data, its dataset-specific frequency was computed. Then, for each word, the PS set frequency was divided by the neutral set frequency (giving the PS-to-neutral ratio), and the NS set frequency was divided by the neutral set frequency (giving the NS-to-neutral ratio). The difference between PS-to-neutral and NS-to-neutral ratios were compared for all words in order to find words that are characteristic of the PS and NS data. More details are in 3.1.1, and the results are in Section 4.1.

2.6 Limitations and Potential Problems

2.6.1 The Confidence in and Quality of the Labels

Since nobody has access to anyone else's beliefs except through their expressions of belief, it is impossible to identify anyone as PS (as defined in Section 1.3) with certainty, since they may have a non-science-based belief, and they may think it is rational, even if they haven't expressed it. This is an issue for any attempt to operationalize the PS vs. NS distinction. The choice to discard conflicted users, rather than allow the NS label to override the PS label (since, according to Section 1.3, all it takes is one non-science-based belief to make someone NS), did not seem like a momentous decision, because so few conflicted users were found (16 total). This choice may not have been the right one, but the motivation was the idea was that people who do not take the side of science on any controversial issue are better examples of NS users.

A bigger issue might be the highly subjective nature of the labels, and the fact that they are likely to be significantly dependent on when the label was craeated (since, generally, only the most recent tweets are read or skimmed in three minutes, and the most recent tweets are ever-changing). It seems that a user might be labeled quite differently if labeled twice a few months apart. In fact no quality control experiments in which users were labeled multiple times (or by multiple annotators) to compare labels were done. It is difficult to say *a priori* if reliably identifying someone's perspective based on three minutes of tweet-skimming is possible, or to what extent it is. However it does not seem a wholly unreasonable approach, given that it is probably much faster than more thorough approaches.

2.6.2 Confounding Variables

Knowing from the start that there could be educational, regional, political, and other correlations with the PS and NS perspectives, I made great conscious effort not to use those correlations to find users or to influence data collection in any way. For instance, knowing that there is a correlation between conservatism and global warming denial (Lewandowsky *et al.*, 2013), it was necessary to self-consciously avoid taking advantage of the fact that someone who seems conservative is more likely to to have a non-science-based idea about climate change. For example, if you search Twitter for “global warming”, and there are multiple result tweets that do not clearly express global warming denial, it may still be worth further investigating these users, so find if they have other tweets that clarify their position. However, if you only choose one user to further investigate, the choice must not be influenced by inferences based on the user’s apparent political ideology. However, given that such correlations have been found to exist, it is not problematic, or necessarily undesirable, if they ultimately are reflected in the dataset: the important thing is not to use such correlations to find users.

Chapter 3

Approach

The purpose of the present work is to learn about the linguistic differences in the language used by PS and NS Twitter users in addition to using such differences to automatically identify the perspective of users.

3.1 Data Analysis

3.1.1 Relative Frequency Ratios

In order to examine word choice differences between the two groups, word frequencies are calculated for the two main datasets (PS and NS user datasets). Comparing these raw frequencies to one another is less informative than using a neutral baseline frequency to compute frequency ratios. In addition to the tweet data carrying PS and NS labels, a third dataset, intended to serve as a neutral dataset, was made by streaming a random sample of tweets (filtering out non-English tweets in the sample). The ratios of word frequencies of the PS data, and those of the NS data, to the neutral data word frequencies are calculated in order to get a sense of to what degree the two non-neutral groups' language differs from "neutral" Twitter language.

3.1.2 Topic Modeling

Topic modeling can be useful for summarizing collections of documents and for summarizing documents. Rather than representing a document as a point in high-

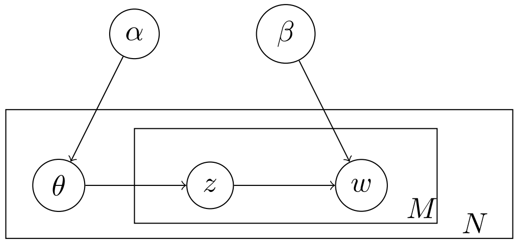


Figure 3.1: Latent Dirichlet allocation

dimensional space, where the dimensions are all the words of the vocabulary, a document can be represented as primarily a mixture of small number of topics[1]. *Latent Dirichlet Allocation* (LDA) (Blei *et al.*, 2003) is an unsupervised generative topic model. It is represented by the graphical model given in Figure 3.1, in which α is the Dirichlet prior on the topic distributions of the document; β is the Dirichlet prior on the distribution over words in the topic; θ is the topic distribution; z is a topic; w is a word; M is the number of documents; N is the number of words in the document.

Building an LDA model from a document corpus can reveal topics in text. These topics may or may not correspond to human judgments about what constitutes a "real" topic. To a human, the idea of a topic is intuitively quite clear, if not necessarily easy to define precisely. But for LDA, a "topic" is nothing more than a probability distribution over the (corpus) vocabulary. It is common to find a variety in the human-judged quality of topics built using LDA.

Aligning Topics from Different Datasets

Intuitively, topics exist independently of perspectives, so that there can be discussion of a topic from different perspectives. In order to study different perspectives, or topics, one might try to look at how the same topics are talked about from different perspectives. Ahmed and Xing (2010) present a new generative model based on LDA that they use to do this kind of analysis. Their model combines the unsupervised topic discovery of LDA with the perspective labels on each document in the corpus to automatically visualize topics from the two perspectives. Lin *et al.* (2008) also separately model topic and perspective. Their model involves giving words both

[1] e.g. 75% topic A, 23% topic B, and 2% other topics

topic weights and perspective weights. In both cases, the model, based on weights or probabilities, can reveal differences in the way topics are discussed from different perspectives. Another approach to exploring how people with different perspectives talk about the same topic involves after-the-fact topic *alignment*. Chuang *et al.* (2013) find that the success of such alignments can be greatly affected by small changes in parameterizations of the topic models.

As an avenue of exploration into the similarities and differences between the PS and NS language, I examine the way topics are discussed from different perspectives. Unlike Lin *et al.* (2008) and Ahmed and Xing (2010), I do not use a generative model that models topics and perspectives in documents. Instead, I take a simpler approach: topics from LDA models built separately from the two datasets are aligned or matched to make "meta-topics". To build these models, MALLET is used (McCallum, 2002). Since an early goal of this research was to see how people with different ideologies and attitudes write about the same underlying topics, it seemed appropriate to see how successful such topic matching would be. One could imagine a situation where many of the topics match up nicely, as a result of my effort to find people from both perspectives talking about the same topics. However, there are significant imbalances in the corpus. For instance, fewer than ten PS writers are labeled with the "astrology" topic.

Matching topics (i.e. distributions over words) involves measuring the similarity between distributions. Although there are many different metrics for that, *Jensen-Shannon divergence*, given in Equation 3.1 is an appropriate one. In the context of matching topics from different perspectives, one virtue is that, unlike *Kullback-Leibler divergence*, given in Equation 3.3, it is symmetric.

$$JSD(P \,||\, Q) = \frac{1}{2}D_{KL}(P \,||\, M) + \frac{1}{2}D_{KL}(Q \,||\, M) \tag{3.1}$$

where

$$M = \frac{1}{2}(P + Q) \tag{3.2}$$

and, D_{KL} is the *Kullback-Leibler divergence*, defined by

$$D_{KL}(P \,||\, Q) = \sum_{i=1}^{n} \ln\left(\frac{P(i)}{Q(i)}\right) P(i) \tag{3.3}$$

Word distributions can only be compared if they are distributions over the same words, so the PS distribution had to be smoothed using the words that only appeared in the NS distribution, and vice versa. They were smoothed by the same amount that inter-model topics are smoothed by in MALLET's implementation of LDA.

I treat the alignment or matching of topics from the PS to ones from the NS data as bipartite graph matching, or more specifically a version of "the assignment problem" (Kuhn, 2005). In the assignment problem, one must assign workers to tasks, where assigning any given worker to any given task has associated with it a non-negative numerical performance score, where the goal is to assign workers to tasks such that the sum of the performance scores of all assignments is maximized. Alternatively, and perhaps more commonly, instead of maximizing performance scores, the goal is to minimize costs, which, like performance scores, are non-negative numbers associated with assigning workers to tasks. In the context of topic matching, the assignment problem would be one of cost minimization. The "cost" of any potential assignment is the Jensen-Shannon divergence between the topics; this metric was chosen over Kullback-Leibler divergence because the latter is symmetric. Jensen-Shannon divergence is defined in Equation 3.1. The assignment problem can be construed as bipartite graph matching.

3.2 Classification

3.2.1 Perspective and Ideology

There is a successful tradition of computationally modeling perspective in text. As far as I am aware, Lin *et al.* (2006) were the first to attempt to automatically detect the perspective of a documents, where the task was to identify whether a document from the BitterLemons corpus was from the Israeli or Palestinian perspective. Based on the classification results, the authors say that "much of a documents perspective is expressed in word usage, and statistical learning algorithms such as SVM and naïve Bayes models can successfully uncover the word patterns that reflect author perspective with high accuracy". In their study, the only features used for classification were unigrams. They tried various models, including naïve Bayes (NB) and support vector machines (SVM), and found that the results were "com-

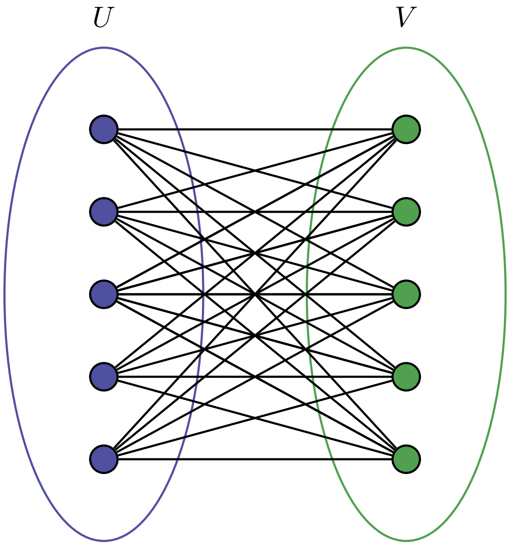


Figure 3.2: Example graph of the first stage of topic matching for two sets of LDA topics. Nodes are topics, U is the set of topics discovered in dataset A, and V is the set of topics discovered in dataset B. Edges are weighted with the Jensen-Shannon divergence between topics. An optimal one-to-one matching, based on edge weights, is acheived quickly ($O(n^3)$) with the Hungarian algorithm..

parable". In all of their models, frequency information is used, rather than binary presence/absence features.

Klebanov *et al.* (2010)'s work is in many ways an expansion of Lin *et al.* (2006) in which classification experiments are performed on datasets made from BitterLemons, BitterLemons-International, the University of Maryland Death Penalty Corpus, and transcripts of the House and Senate floor debates on Partial Birth Abortion Ban Act. They use some of the same models as Lin *et al.* (2006), in addition to others. Namely, models are tested that use binary presence/absence features) and ignore frequency. They fail to find evidence that frequency information improves perspective classification performance over presence/absence information on any of the datasets. They find that for all tasks, using an SVM model that utilizes binary presence/absence features performs the best, or not significantly differently from the best. Thamrongrattanarit *et al.* (2013) have success detecting restaurant reviews from the vegetarian perspective using distant supervision and multinomial NB. Regarding the BitterLemons corpus, Hardisty *et al.* (2010) established a new state-of-the-art for the data set. They present a non-parametric version of NB, in which specific n-grams of large n are used if they are discovered automatically to be good features, so that, for example, a 6-gram like "get the government out of my" might be used without using other 6-grams. This solves the sparsity problem of using larger n-grams.

Related to supervised perspective classification is the effort to automatically identify perspectives, either using another source of knowledge, as Gottipati *et al.* (2013) does with Debatepedia, or in a way that requires no outside source and no supervision, but only the guarantee of properly paired (i.e., topically similar) datasets, as done by Paul *et al.* (2010). I do not venture into this territory in this project; the hope is that the quality of classification based on fully supervised learning is superior to the quality that would be achieved otherwise.

3.2.2 Perspective (and Lack Thereof) on a Sub-Document Level

It was not obvious, initially, that the concentration of tweets that revealed a perspective would be sufficient for achieving good classification results. It seemed like articles would be an easier task, since Twitter users, even if they are advocates of science or some kind of non-science (e.g. pseudoscience), often tweet about current events, TV shows, sports, etc, between their PS or NS tweets. Others in this area of research had similar concerns when they approached their tasks. Sim *et al.* (2013) use the term "ideology", but what they are doing could easily be called perspective modeling. Concerned with political ideology, they depart from binary classification, and attempt to model political position as a mixture of hierarchically related ideological/political categories (e.g. "libertarian", a subcategory of "right"). These categories also correspond to states in an *hidden Markov model* (HMM) that generate terms ("cues") characteristic of that category; in their experiment, they learn the cues for categories from data. Between cue words, in their model, are "filler words", i.e. any words that their algorithm did not select as a perspective-revealing cue word. The idea that there really are only a few perspective-revealing "cue" words intuitively seems right.

In doing classification on the BitterLemons corpus, Lin *et al.* (2006) are concerned with the fact that many sentences of a document do not reveal a perspective, even if the document overall does, noting that "when an issue is discussed from different perspectives, not every sentence strongly reflects the perspective of the author." They present a generative model based on naïve Bayes (NB), where perspective and sentence perspective are both modeled. In their model, although every document must have perspective, every sentence does not. On two classification tasks, its accuracy is "comparable to or even slightly better than that of" NB. Since they are not using a corpus with sentence-level gold labels, they cannot truly evaluate the sentence-level modeling, but they argue their model's performance at the document level suggests that the sentence-level modeling is working reasonably well. Given that their results are extremely similar to simply using unigrams and NB, I chose to go the simpler route by which I made no effort to model the perspective or lack thereof for individual sentences or phrases within the document.

3.2.3 Twitter Users

Rao and Yarowsky (2010) successfully use stacked-SVM-based classification algorithms to classify users according to various latent attributes, including gender, age, regional origin and political orientation. the classification done in my thesis falls under the heading of latent attribute classification. So the work of both Rao and Yarowsky (2010) and Burger *et al.* (2013) are related to my classification tasks.

Their success suggests that classifying Twitter users by PS and NS perspectives is a reasonable task. While they experiment with using features based on information other than the language of the tweet text itself to classify users, I avoid that, because I am interested in the way that perspective is reflected in normal text, rather than any other kind of network information or metadata associated with tweets or Twitter accounts.

3.2.4 Distant Supervision

Thamrongrattanarit *et al.* (2013) get quite good results even though their training data is not labeled by an annotator. I attempt something similar using Twitter's user-curated Lists, except I take steps to guarantee high precision, as they do. They filter reviews searching for a few specific phrases that virtually guarantee that the review is from a vegetarian perspective, including "as a vegetarian" and "I'm a vegetarian". This method, as a detector for reviews written by vegetarians, increases precision at the expense of recall. Since my method involves no per-tweet, or even per-user filtration, precision likely suffers.

3.2.5 Logistic Regression

Various types of classifiers would be reasonable choices for my task. I chose logistic regression because I conjectured that, as a discriminative model, it would be approximately as good or better than generative classifiers, such as NB, or the model inspired by NB that Lin *et al.* (2006) present. My choice of logistic regression over other discriminative classifiers, e.g. support vector machines, was arbitrary. I use LIBLINEAR's L2-regularized logistic regression.[2]

One basic difference between this classifier, and NB classifiers, is that, unlike logistic regression, the NB model weighs each feature independently. This may sound undesirable, but NB does not always underperform against logistic regression. It appears that NB may be superior when there is less training data (Ng and Jordan, 2002). Lin *et al.* (2006) get their best results with their NB-based model, rather than SVM, and they similarly speculate it involves the fairly low number of training

[2] LIBLINEAR is a software library for building discriminative classifiers. It is written in C++. In my code, I use the Nak machine learning library (`https://github.com/scalanlp/nak`), which in turn uses a Java port of LIBLINEAR.

documents (their corpus is 594 articles). Wang and Manning (2012) show that both NB and SVM can achieve state-of-the-art classification results: the key is in featurization. One of their findings is that for longer documents, SVMs outperform NB in sentiment analysis tasks. It may be that this would be the case of perspective classification too. However, in this thesis logistic regression is used for all tasks.

3.2.6 Features

Various ways of featurizing documents are evaluated, including unigrams, hashtag-based features, character n-grams, and *Linguistic Inquiry and Word Count* (LIWC) (Tausczik and Pennebaker, 2010) counts. The hashtag-based features were basically used instead of attempting to segment hashtags into individual words, and then making features out of those (hashtags cannot contain spaces, so they are either one word, or they are multiple words strung together with no spaces).Unigram and bigram features are standard for text classification, but LIWC features and the hashtag-based features used here are not.

The core of LIWC is the LIWC dictionary, which is a set of word categories and a specification of exactly what words belong in each category. For instance, there is a "positive emotion" category that includes words like "love" and "sweet". There is hierarchy in the LIWC categories, so different categories may or may not be disjoint. A LIWC analysis details the proportion of a given document's words that fall within each of the LIWC dictionary's categories[3]. The LIWC analysis can be straightforwardly turned into a vector of features for classification (which I call "LIWC features"). These features have occasionally been used in text classification, (e.g., Stark *et al.* (2012) use them in this way), though I am not familiar with a study in which they prove to be particularly useful in text classfication. For instance, Stark *et al.* (2012) find that using LDA topic-based features give superior results for classifying phone conversations by their social nature. However, given the psychological validity of LIWC features (Tausczik and Pennebaker, 2010), and the apparent differences in the way PS and NS think (Pennycook *et al.*, 2012), the idea that they might prove useful for my classification task seemed plausible.

Hashtag-based features were used in lieu of attempting to segment hashtags

[3] e.g., a LIWC analysis of a document containing only the word "love" would indicate, among other things, that the proportion of words in the document falling into the "positive emotion" category is 1.0.

into words. When hashtag features were used, one or multiple kinds of features were created from each hashtag. If the hashtag was fewer than four letters long, a “short hashtag” feature was made[4]. If the hashtag was four letters long, the hashtag was featurized as a 4-gram[5]. If the hashtag was five letters or longer, it was made into all possible 4- and 5-grams[6].

In order to keep things simple, for all classification tasks, all query terms ever used for finding users were filtered out for unigram and LIWC featurization, and all of them were replaced with the string “[-queryword-]” prior to bigram featurization. This was necessary to avoid corrupting classification results. After all, different query terms were often used in attempts to find users of different perspectives; using those very terms as features to classify documents could make my classification tasks self-fulfilling.

[4] shortHT=⟨full hashtag⟩

[5] 4gram=⟨full hashtag, i.e. the only 4 gram⟩

[6] e.g. for a five-letter hashtag: 4gram=⟨first 4-gram⟩, 4gram=⟨second 4-gram⟩, 5gram=⟨full hashtag, i.e. the only 5 gram⟩

Chapter 4

Analysis

The two perspectives were compared in terms of word frequency (using frequencies relative to frequencies of the neutral tweet dataset) and in terms of topics generated discovered with LDA.

4.1 Relative Frequency Ratios

Table 4.1 and Table 4.2 (which is a continuation of Table 4.1) give a list of words, ranked by how different their frequency ratios (relative to the neutral data) are. Smoothing was used to ensure that no word count in any of the three datasets was zero (to avoid the possibility of a denominator being zero in calculating frequency ratios)[1]. The "NS", "PS" and "neut." columns give raw counts. Italicization was used to make interpreting the list easier: words in italics are the ones that are more common in the PS data, wheras unitalicized words are more common in the NS data. Asterisks indicate that the word was used as a query or part of a query in data collection (see Appendix A for the full list of queries), so the word's high ranking is potentially only due to its use in queries). For this list, only words that appeared twice in all three datasets are included. Furthermore, non-words (where the "word" includes symbols or digits) were filtered out of this list. Proper nouns were not filtered out.

[1] Smoothing went as follows: if any word in the set of words contained in any of the three datasets was not in a particular one of the three datasets, rather than using 0 as the count of that word in that dataset, for the purposes of the calculation, 1 was used for the count

word	**diff**	**PS/neut.**	**NS/neut.**
compatibility	1035.4	3.0	1038.5
retrograde	500.4	4.2	504.6
shui	359.9	5.6	365.5
scopes	317.0	5.6	322.6
astrology	212.2	12.1	224.3
atheism*	205.2	217.2	12.0
reiki	204.0	5.9	209.9
astrological	176.4	2.3	178.8
humanist	165.2	171.3	6.1
audiobooks	156.7	161.5	4.8
vaccination*	153.9	195.8	41.9
measles	151.7	163.9	12.3
feng	150.3	2.8	153.1
saturn	149.5	17.6	167.1
atheist*	146.8	155.1	8.4
creationist	146.7	148.0	1.4
atheists	145.7	160.2	14.4
horoscope	143.5	0.4	143.9
moon's	143.0	11.2	154.3
anti-vaccine	132.1	135.2	3.1
agnostic	123.3	136.2	12.9
jal	123.0	0.5	123.4
creationism*	112.9	115.8	2.9
skeptics	108.0	121.9	13.9
vaccinated	107.0	202.9	96.0
ufo*	103.5	2.6	106.1
cpac	91.5	97.4	5.9
humanism	87.3	94.1	6.8
evolutionary	85.8	132.7	47.0
mercury	84.0	6.7	90.6

Table 4.1: Words ranked by the absolute value of the difference between PS frequency over neutral set frequency (given in the **PS/neut.** column) and NS frequency over neutral set frequency (given in the **NS/neut.** column). Non-words (where the word includes symbols or digits) were filtered out of this list. First 30 words. Continued in Table 4.2.

word	**diff**	**PS/neut.**	**NS/neut.**
hpv	77.0	142.3	65.3
superstition	76.3	82.3	6.0
fullness	74.1	2.8	76.9
manifesting	73.9	2.8	76.7
vaccinate*	71.4	89.9	18.5
intellect	69.7	4.8	74.5
dissonance	69.2	82.2	12.9
vaccinations	67.7	120.8	53.1
skeptic	67.6	75.0	7.3
hitchens	67.2	68.8	1.6
affirmations	65.6	2.1	67.7
sep	63.8	5.9	69.7
scathing	63.6	65.7	2.0
deniers	63.2	64.4	1.2
indoctrination	62.2	66.6	4.4
denier	61.9	63.9	2.0
visualizing	61.5	6.1	67.6
voc	61.2	3.3	64.4
affirmation	60.8	3.8	64.6
prophetic	60.0	5.6	65.6
planetary	59.8	47.8	107.5
uranus	59.0	3.1	62.1
dishonesty	58.7	70.9	12.3
mutilation	58.5	66.0	7.5
disprove	58.5	66.0	7.5
kepler	58.5	63.9	5.4
chiropractors	57.8	63.2	5.4
vaccines	56.2	197.4	141.2
mccarthy	54.5	59.6	5.1
tarot	51.0	1.2	52.3

Table 4.2: Words ranked by the absolute value of the difference between PS frequency over neutral set frequency and NS frequency over neutral set frequency. Non-words (where the word includes symbols or digits) were filtered out of this list. Continued from Table 4.1. Words 31 – 60.

The results presented in the tables often correspond to what would be expected by someone familiar with the perspectives and topics. For instance, my impression from data annotation is that the term“anti-vaccine” is associated with the PS perspective; people who are actually opposed to vaccination, when referring to their position are more likely to describe themselves as being in favor of “vaccine choice”. Similarly, “retrograde” is an astrological term, but is quite a rare word otherwise, and, in general, I found that PS users rarely use astrological terminology, even when stating their disbelief in astrology. Few PS users discussed the actual “inner workings” of astrology. Though not clear in the table , this seems different from some other topics, such as monotheistic religion topics, where PS users often mention the divine characters by name, even citing bibilical material in the tweets in which they express their disbelief. In any case, the words in Tables 4.1 and 4.2 provide insight into the differences between the words people with each perspective use.

4.2 Topic Modeling and Alignment

Topic models for the PS and NS data were separately made using MALLET (McCallum, 2002). In both, $K = 30$. This number was arbitrarily chosen. In 3.1.2, I explain the process of smoothing the topics from each model with vocabulary from the other model, in order to have distributions over the same vocabulary, which is a prerequisite for calculating Jensen-Shannon divergence. The matchings are represented as side-by-side top 25 word lists in tables. Some noteworthy matches are given and discussed here.

The quality of topics and alignments are judged by looking at the top words (no other attempt to evaluate the LDA models is used). Some of the alignments are good in my judgment. Out of 30 matches, 12 were at least reasonable, but only 9 were very good. Of the 12 reasonable matches, 5 are relevant to this thesis.

Two of the matches form metatopics under the heading “religion”. One of them is in Table 4.3. It seems to represent a Christian subtopic of religion, with the exception of the word “Islam” on the PS side. Top words on both sides include God, Jesus, and prayer. The LDA models provide Dirichlet parameters for topics; in this case, 0.30792 for PS and 0.19859 for NS. Within an LDA model, these parameters roughly indicate the weight of the topic. Since we are comparing topics from two models, the Dirichlet parameter is divided by the sum of all Dirichlet parameters from the relevant model. The sum for PS is 13.37, and for NS it is 5.57. So instead of considering the previously mentioned parameters, one should rather consider the weight for the PS topic to be 0.023, and that of the NS topic to be 0.036. This could be taken to suggest that religion is a more prominent topic in the NS data, which is accurate according to Table **??**. But this approach could have problems as a way to tell which underlying topics are weighted more in which dataset. It is more reasonable to sum all the religion-like topics from both models, matched or not, to get a better indication of the true weight of the religion topic in the two datasets. This idea will be revisited after discussion of other notable results of the matching.

The other "religion" metatopic is in Table 4.4. This metatopic seems to be matching a more tradition religion topic on the PS side with a more spiritual one on the NS side. This matching makes sense to me, because spirituality seemed a more common topic among NS users. In my experience, PS users usually focused on popular monotheistic religions rather than other kinds of spirituality.

Table 4.5 has a few religious terms on the PS side, but overall it seems to be a metatopic of evidence/reasoning/truth. This is especially interesting for me because I didn't know there was such a topic within the NS data. During user annotation, such a topic was observed, but rarely enough that it was never identified using a topic or subtopic label (subtopic labels are not used in any way in this thesis, but they were given to most users). The only query that would obviously help account for topics resembling these was"soul AND scientific AND evidence". But only one user was labeled as a result of using that query. This matching may be the most intriguing one.

Another good match is in Table 4.6. However, this is not necessarily an ideal meta-topic, at least based on the top 25 words: the fact that the aligned topics are from corpora of different perspectives on science does not really come through, or if it does, it is to a lesser degree than for the previously mentioned matches. One would have difficulty guessing which topic came from which dataset. This may be attributable to the limitations of LDA (which does not puport to capture perpsective in topics).

Not surpisingly, many topics from both models seem poor. In some instances a poor LDA topic seems like a plausible cluster, based on intuitions about co-occurrence, just not a "real" topic (in the normal, non-LDA sense of the word

PS	NS
god	god
religion	jesus
atheist	lord
atheists	christ
religious	life
jesus	love
bible	faith
christian	grace
church	word
people	spirit
christians	prayer
faith	heart
love	man
world	sin
atheism	world
day	people
islam	good
hell	give
good	pray
happy	holy
life	things
book	today
ham	father
death	amen
prayer	bible

Table 4.3: Matched topics. Top 25 words. JSD: 0.375

PS	NS
religion	meditation
godless	spiritual
sunday	eye
religious	mind
life	jal
back	knowledge
human	body
man	nature
church	light
world	enlightenment
existence	spirit
society	consciousness
philosophy	love
thinking	life
mind	higher
street	ego
fiction	namaste
belief	human
word	soul
robert	energy
mankind	beings
proud	practice
free	yoga
ideas	existence
minds	positive

Table 4.4: Matched topics. Top 25 words. JSD: 0.535

"topic"; in some situations, to avoid ambiguity, I will use the term "LDA topic"). An example of this might be an LDA topic where the top words are curse words. Other times a topic may not seem like a plausible cluster, i.e. where it is not even obvious that the top words would co-occur. Such topics may simply be junk, or they may be an instance of *fusion* of topics. Chuang *et al.* (2013) try to study the effects of different parameterizations of LDA models in order to minimize both junk topics and fusion of topics, while trying to increase the number of "resolved" topics when doing topic alignment. "Resolved" topics are real topics that are successfully represented as an LDA topic in the model (the terminology is again from Chuang *et al.* (2013)).

PS	NS
god	people
evidence	truth
evolution	read
science	world
true	human
bible	science
claim	real
make	true
people	wrong
wrong	children
religion	years
belief	evidence
exist	fact
understand	point
fact	question
faith	earth
read	called
proof	history
prove	man
question	understand
universe	based
gods	problem
exists	death
man	made
answer	change

Table 4.5: Matched topics. Top words. JSD: 0.347

PS-8	NS-8
health	health
study	autism
risk	food
cancer	vaccine
flu	study
patients	cancer
disease	medicine
medical	body
care	vaccines
medicine	natural
drug	children
good	healthy
heart	homeopathy
food	dr
doctors	safe
healthy	diet
high	flu
vaccine	medical
doctor	news
weight	raw
diet	disease
blood	foods
kids	free
brain	gmo
fda	drugs

Table 4.6: Matched topics. Top words: JSD: 0.273

Alignments can seem poor either because one or both of the aligned topics themselves have problems described above, or because they appear misaligned.

The simple matching technique used here is easy and reveals some things about the similarities and differences in the datasets, but has fundamental weakness that most likely prevent it from being anything other than an exploratory technique. The biggest weakness may be the insistence that every topic from each dataset be matched. This most likely results in sacrifices being made wherein the best matches are not always chosen. Perhaps using an algorithm for pruning topics that are highly divergent from every topic in the opposing set could imporve results.

It is worth finishing the discussion of what the Dirichlet parameters indicate about the prominence of topics in the data. By my judgment, there are two religion/spirituality topics in the PS set of topics, and three in the NS set. Summing the Dirichlet parameters and dividing by the topic-wide total, one finds an overall religion/spirituality weight of 0.035 for PS and 0.078 for NS; and for health, one finds the weight to be 0.037 for PS and 0.018 for NS. This generally corresponds with the observation that health was a lot more prevalent in the PS data and that religion was more prominent in the NS data (see Table 2.7). There are four astrology-related topics on the NS side, but none on the PS side, so there is no need to talk about parameters here. This imbalance is not particularly surprising given the imbalance seen regarding the topic of astrology during the user annotation process (20 users in PS; 169 in NS).

Chapter 5

Classification

The classification tasks were to classify Twitter users and to classify articles.

5.1 Development: Selection of C and Features

To develop the best possible classifier, various values for C, which is a penalty parameter (see Section 3.2.5), and various approaches to featurization were evaluated by obtaining classification accuracy on the dev set. Every value for C that was used that was higher than 0.5 resulted in the same classification results. A value of 1.0 was arbitrarily chosen from that range to parameterize the final classifiers. The best classification results came from using presence/absence unigram features and presence/absence hashtag-based features. As a baseline for classification, the most common label in the training set is used as the prediction for every document in the evaluation set. Unless indicated otherwise (e.g., UNI(FREQUENCY)), for all of the featurization types, presence/absence features were used, rather than features that took frequency into account, except for LIWC features, which did encode frequency. A few featurizations, and the resulting accuracy, are given in Table 5.1. Note that HT means "hashtag-based features" in the Tables in this chapter. Note that user mentions and occurences of "RT" and "MT" were stripped out of tweet text prior to featurization; many emoji, punctuation-based smileys, and URLs[1] were

[1] All URLs identified as such in the tweet metadata from the Twitter API were removed; however, URLs that were not identified as such by Twitter were not removed. The unidentified URLs correspond to non-hyperlink URLs in tweets.

features	**accuracy**
BASELINE	50.5
UNI	92.9
UNI(STEMMED)	92.4
UNI(FREQUENCY)	89.9
UNI(STEMMED + FREQUENCY)	90.9
BI	91.4
BI(STEMMED)	91.4
LIWC	64.7
HT	87.9
UNI+BI	91.4
UNI+HT	**94.4**
UNI+LIWC	90.4
UNI+HT+LIWC	91.9
UNI+BI+HT+LIWC	90.9

Table 5.1: Accuracy on the dev set using various featurizations

also removed prior to featurization.

5.2 Main Classification Task

Using the best C and featurization found when testing on the dev set, an accuracy of 93.9% was attained on the test set. The small drop from the accuracy of 94.4% on the dev set is not surprising. Table 5.2 is the confusion matrix. Misclassifications were almost evenly split: 5 PS users were misclassified as NS, and 7 NS users were misclassified as PS.

A similar classification experiment was done, in which all automatic retweets (unmodified retweets) were stripped out of both testing and training data. The accuracy was identical, at 93.9%, suggesting the issue of retweeting is ultimately unproblematic.

5.3 Training and Testing on Different Numbers of Tweets

The average number of tweets in each document was 1084. However, one could imagine using fewer tweets and getting similar accuracies. In order to test this

		predicted label		
		PS	**NS**	**total**
actual label	**PS′**	93	5	
	NS′	7	93	
	total	P	N	

Table 5.2: Confusion matrix for testing on the test set. Overall accuracy: 93.9

possibility, I performed experiments where the number of tweets actually used from each evaluation document varied. Figure 5.1 shows accuracy as a function of the maximum number of tweets being used[2] for each document in the evaluation set; the entirety of the documents are still being used in training. Often, using thousands of a user's tweets to categorize them is not practical. The graph shows accuracy leveling off at 93.9% at about 500 tweets, equivalent to the accuracy attained when the entirety of the tweets in each test document are used). This supports the idea that using about 500 tweets is sufficient for high accuracy categorization of PS and NS.

5.4 Training and Testing on Different Label Topics

One question I had when I started this project is whether the PS and NS perspectives cut across topics enough that someone with an NS attitude about a topic that rarely appeared in the training data could still be identified using my approach. For this experiment, I tried every combination of topics for training and testing. The results

[2] It is the maximum in the sense that, if the user document contained x or more tweets, then x tweets were used, but if the user had fewere than x tweets, than all of the user's tweets were used. For instance, when 1,000 is used as the maximum number of tweets, only 102 tweets are being used for the user that only has 102 tweets.

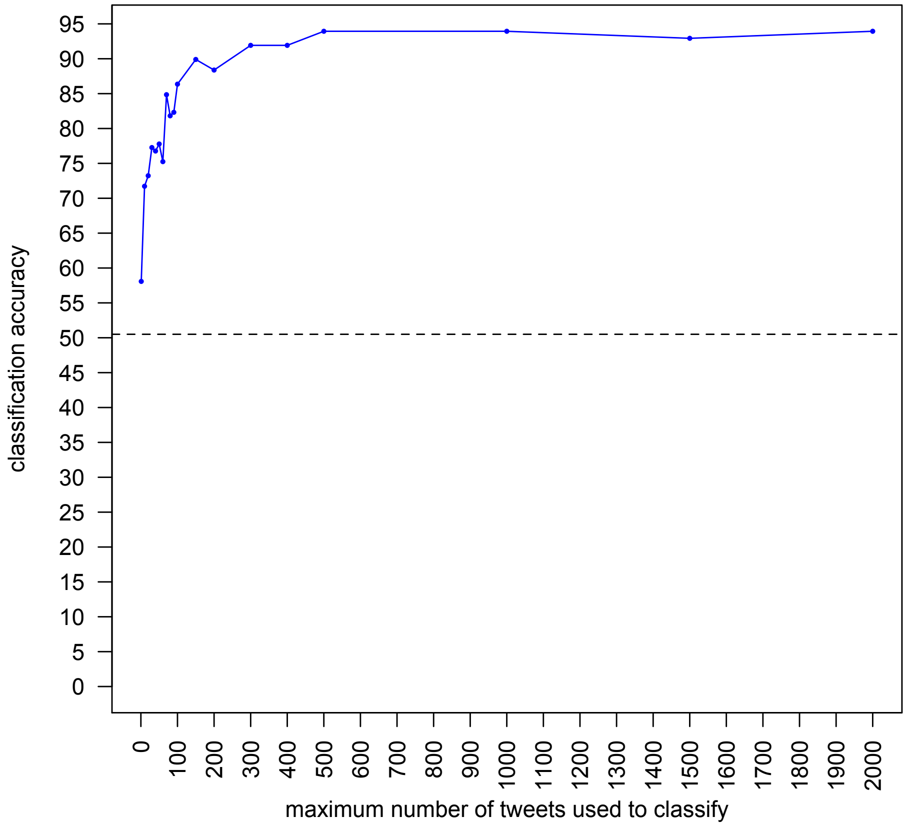


Figure 5.1: Using the test set, training on the full training user tweet documents; evaluating on varying number of tweets. Most-common-label baseline accuracy is 50.5%, and it is shown as a dashed line in the graph.

		Test		
		Health	Religion	Paranormal
Train	Health	84.0	89.0	64.0
	Religion	69.1	100.0	88.9
	Paranormal	55.6	80.0	88.9

Table 5.3: Dev set: Testing and training on various topics. N.B. some users are in more than one of the three topic categories, and are in multiple categories; training and testing on "different" topics does not guarantee their topics are disjoint.

		Test		
		Health	Religion	Paranormal
Train	Health	87.8	83.8	70.3
	Religion	80.4	97.0	89.2
	Paranormal	64.6	75.8	86.5

Table 5.4: Test set: Testing and training on various topics.

for the dev set are in Table 5.3, and the results on the test set are in Table 5.4. It is important to understand that, for this experiment, each by-topic dataset is made up of users with that topic label, but not exclusively that topic label. For instance, someone in the "health" training set may also have a "relig" label, and someone in the "relig" test set may also have a "health" label. Therefore, these experimental accuracies may be inflated by a small amount of overlapping topical content for different topic subsets.

5.5 Testing on Different Levels of Extremity of Perspective

To see if extremity of belief had a big influence on the probability of correctly classifying a user, in one experiment, the best classifier (trained on the full training data using unigrams and hashtag-based features) is tested on subsets of users in the test set that are divided based on level of extremity that the user appears to express. The results, which show a fairly large effect, are in Table 5.5.

testing on	ACCURACY
LOWER EXTREMITY	86.2
HIGHER EXTREMITY	95.1
HIGHEST(ADVOCATES)	95.8

Table 5.5: Accuracy when testing on subsets based on extremity of perspective.

testing on	ACCURACY
RARE	88.9
LOWER	91.7
HIGHER	94.9

Table 5.6: Accuracy when testing on different frequency-based subsets. The test set is divided into "lower" and "higher" frequency. "Rare" is a subset of "lower frequency."

5.6 Testing on Different Frequency of Belief Expression

In order to see if the frequency of perspective-revealing tweets affected accuracy of classification, the test set was divided up by their frequency label. The results are in Table 5.6. As with extremity of beliefs, frequency has a noticeable effect on accuracy. It is interesting that accuracy is still respectable when the frequency is rare. For some of these users, only a single tweet was found that revealed the perspective with which they were labeled.

5.7 Training on Noisy List-based Dataset

There were two tasks where the classifier was trained on the List-based (i.e., noisy) data. In the first, the number of total number of users is held to 922, to match the size, in terms of documents, of the individually-labeled training set. The idea here is to see if there is a decrease in accuracy attributable to the fact that the labels are noisy. In the other task, many more documents are included, just under three times as many as in the smaller (922) set, in order to see if this increases accuracy. The closer the accuracy of the "List-trained" classifier is to one trained on individually-labeled data, the less reason there is to ever use the time and resources to individually

	UNI+HT	BASELINE
922 DOCUMENTS	81.8	49.5
2,693 DOCUMENTS	82.8	49.5

Table 5.7: Accuracy when training on noisy data; testing on dev set.

	UNI+HT	BASELINE
922 DOCUMENTS	83.8	50.5
2,693 DOCUMENTS	82.8	50.5
INDIV + 922	90.4	50.5

Table 5.8: Accuracy when training on noisy data or noisy data in addition to the main training set; testing on test set.

label twitter users for this kind of task. This approach would be moving in the direction of distant supervision, somewhat akin to Thamrongrattanarit *et al.* (2013), where a small amount of effort is used to automatically collect a lot of training data. Accuracy for the two experiments is in Tables 5.8 (dev set results) and 5.8 (test set results). The test set table includes an additional result for "INDIV + 922", which means the training data is the combination of the 922 users with noisy labels and the training set of individually labeled users used for the full supervision experiments. The experiment is meant to see if augmenting strongly labeled documents with weakly labled ones can improve PS/NS classification.

The results are considerably worse than the results from training on individually labeled users, but it may be good enough for some applications. Acquiring the noisy data took a tiny fraction of the time and effort required for individually labeling users; in fact, thousands of users can be given noisy labels in the same amount of time that it takes to give one user an individualized label. When individually-labeled and group-labeled users is combined, it seems that the accuracy (90.4%) is decreased relative to excluding group-labeled users from training (which yields 93.9% accuracy). However, this could still be a fast and easy approach for expanding the topic familiarity of the classifier (an idea which is not explored here).

		predicted label		
		PS	**NS**	**total**
actual label	**PS′**	58	41	
	NS′	11	90	
	total	P	N	

Table 5.9: Confusion matrix for testing on the article set. Accuracy is 74.0%.

5.8 Testing on Article Dataset

This experiment gives an indication of how similar or different twitter timelines are from articles as expressions of perspective. Note that, since articles do not use hashtags, those features were not at the disposal of the model for this task. The classifier was trained on the main dataset, using presence/absence unigram features only. Note, this is a different model from the "uni+hashtag" model used for every other final evaluation. Accuracy is 74.0%. The most common label baseline is 50.5%. The confusion matrix is in Table 5.9. The relatively low accuracy suggests that Twitter data is insufficient for making a PS vs. NS classifier for articles.

5.9 Errors

In order to see what kinds of documents give the classifier the most trouble, one can examine the errors when testing on the dev set. There are two observations that jump out. First, 9 out of the 11 misclassified users have only a "health" topic label, and all but one of these 9 have exactly one subtopic label: "vaccines". For only 4 of the misclassified users did the classifier assign a probability below 0.8 of being in the (incorrect) class; i.e., the classifier was quite confident about what turned out to be errors, even assigning a 0.99 probability of the PS label to one NS user. Combined with the data from testing on users with the health label in Tables 5.4 and 5.3, this is strong evidence that users with health labels are the hardest to classify as NS or PS.

5.10 Issues and Possible Criticisms

In all the classification experiments in which retweets were not excluded from the user's tweet document, it is virtually certain that some of the same tweets appeared (within the context of larger documents) in both the training and test set. This may have inflated results. However it is also plausible that any retweets that created overlap were diluted by the rest of the tweets enough to not make a difference. The fact that classification where retweets are filtered out gave the same results as when they are left in suggests that they don't make much difference.

Another possible objection to what I have done is that, though I claimed to be interested only in the linguistic features, I have used non-linguistic features, viz., LIWC and hashtag-based features. It is true that in some sense, the LIWC features are not feature of the language itself, but features in reference to some other body of knowledge. However, LIWC features are still wholly determined by the language, which is what is important to me. I did not want to use, for instance, someone's social network to predict their perspective, because that sheds no light on the connection between their perspective and the language they generate. Regarding hashtag features, the issue is that hashtags seem to skirt the line between language and metadata. Sometimes they are grammatically separate from the rest of the tweet, and really only seem to be metadata-style tag for the tweet[3] and other times they are decidedly part of the language[4]. Since they are often language, and are, in any case, user-generated metadata, I felt it did not detract from the idea that the classifiers that use hashtag-based features are nevertheless language-based classifiers. Furthermore, I have heard people using metadata-style hashtags in spoken language, which suggest that there is no basis anymore for treating them as different from words.

[3] e.g., "This weather is fantastic! #summer"

[4] e.g., "This #summer is going to be good"

Chapter 6

Conclusion

The analyses and classifications results indicate strong linguistic differences between PS and NS language. This was evident in the ratios of word frequencies PS and NS datasets relative to frequency in a neutral dataset (Tables 4.1 and 4.2). While the topic alignments (shown in the tables in Section 4.1) had limited success in pairing what appear to be two versions of the same real underlying topic, they did reveal similarities and differences. Also, matchings aside, the differences in the LDA models for the PS and NS data supported the idea that health topics are more prominent in the PS data, and religion or spirituality topics are more prominent in the NS data, and generally suggestive of divergence between the two datasets.

The classification experiments in Chapter 5 support the theses I enumerated in Section 1.4. The high overall accuracy of the basic classification task wherein documents from the individually labeled documents in the PS and NS datasets are classified as either PS or NS, using unigrams and hashtag character n-grams, suggests that the distinction between the perspectives corresponds to linguistic differences generally. To specifically test generalizability across topics, classifiers were trained and tested on by-topic subsets, and the results, while less impressive than those from training on the full training data, are fairly good in most cases, supporting the idea of topic-independent perspectives. This is also suggests that the best classifier, trained on all the training data, would generalize reasonably well to Twitter users who discuss topics that don't appear in the corpus that I collected at all[1]. The fact that excluding automatic retweets from the training and testing

[1] e.g., the topic of psychoanalysis

made no difference to accuracy strongly suggests that they neither hurt nor help classification. The fact that accuracy is lower on the subset of users who express lower extremity of perspective and users who express their beliefs or perspective less often was to be expected.

Appendix A

Queries

queries
“believe in ghosts”
“don’t vaccinate”
#andrewwakefield
#atheist OR #atheism
#autism AND #vaccine
#dontgetvaccinated
#dontvaccinate
#getvaccinated
#stopAVN
#vaccinateyourkids
#vaccinechoice
#vaccinedamage
#vaccinedanger
#vaccineinjury
#vaccines AND #bigpharma

Table A.1: Queries that were used to find Twitter users. List continues in following tables.

queries (continued)
#vaccinesaredangerous
#vaccinetruth
#vaxfax
#wakefield #fraud
#wakefield #quack
#witchcraft
(gingko OR ginkgo) AND focus
(ginkgo OR gingko OR ginko) AND focus
(ginkgo OR gingko OR ginko) AND quack
Gemini
Leo
UFO AND sightings
aquarius
aries
astrology
autism AND epidemic
autism AND vaccine
bigfoot AND myth
capricorn
concentrate
creation AND science
creationism AND science
echinacea AND proven
echinacea AND pseudoscience
echinacea AND quackery
echinacea AND unproven

Table A.2: Queries that were used to find Twitter users (continued).

queries (continued)
ghosts AND believe
ghosts
gingko AND mind
god AND science
god science
herbal AND #BigPharma
libra
memorize
mind
paranormal
pisces
psychic AND fraud
psychic AND tarot
quackery
religion AND #Science
soul AND energy
soul AND scientific AND evidence
soul AND vibrations
spirit AND energy
spirit AND vibrations
spiritual
spiritual AND balance
spiritual AND energy
vaccine

Table A.3: Queries that were used to find Twitter users (continued).

Appendix B

Article Sources

website	**number of articles**
vaccinedangers.com	1
themattwalshblog.com	1
imperfectspirituality.com	1
divineharmony.org	4
motherjones.com	1
planetwaves.net	3
whale.to	2
starsdanceastrology.blogspot.com	3
blogs.scientificamerican.com	4
scienceblogs.com	1
spiritualityhealth.com	13
phys.org	1
realastrologers.com	6
skepticblog.org	38
tinybuddha.com	1

Table B.1: Websites that articles were collected from. Continues in the next table.

website	**number of articles**
1-800homeopathy.com	4
christianpost.com	13
theness.com	13
thevaccinemachine.blogspot.com	1
momswhovax.blogspot.com	1
antiantivax.flurf.net	1
freethoughtblogs.com	3
inamirrordimly.com	3
autism-watch.org	3
christianity.com	5
washingtonpost.com	1
skeptic.com	1
alise-write.com	18
shotofprevention.com	1
evangelicaloutpost.com	3
blog.hmedicine.com	2
slate.com	9
modernmom.com	1
drhomeo.com	6
openmarket.org	1
homeopathyzone.com	1
deeperstory.com	2
discovermagazine.com	1
tlc.howstuffworks.com	1
thrivenaturopathicmedicine.com	1
huffingtonpost.com	1
sciencebasedmedicine.org	22
blog.mommeetmom.com	1

Table B.2: Websites that articles were collected from. Continued from previous table.

Bibliography

Ahmed, A. and Xing, E. P. (2010). Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, **100**(3), 407 – 425.

Benson, H., Dusek, J. A., Sherwood, J. B., Lam, P., Bethea, C. F., Carpenter, W., Levitsky, S., Hill, P. C., Clem, D. W., Jr, Jain, M. K., Drumel, D., Kopecky, S. L., Mueller, P. S., Marek, D., Rollins, S., and Hibberd, P. L. (2006). Study of the therapeutic effects of intercessory prayer (step) in cardiac bypass patients: A multicenter randomized trial of uncertainty and certainty of receiving intercessory prayer. *American Heart Journal*, **151**(4), 934–942.

Blackmore, S. and Moore, R. (2014). Seeing things: Visual recognition and belief in the paranormal. *European Journal of Parapsychology*, **10**(1994), 91–103.

Blanshard, B. (1974). *Reason and Belief*. George Allen and Unwin, London.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, **3**(4-5), 993–1022.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2013). Discriminating gender on twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309.

Chuang, J., Gupta, S., Manning, C. D., and Heer, J. (2013). Topic model diagnostics: Assessing domain relevance via topical alignment. In *Proceedings of the 30th International Conference on Machine Learning*.

DeStefano, F., Price, C. S., and Weintraub, E. S. (2013). Increasing exposure to antibody-stimulating proteins and polysaccharides in vaccines is not associated with risk of autism. *The Journal of Pediatrics*, **163**(2), 561 – 567.

Dunning, B. (2012). How to tell a good website from a crap website. `http://skeptoid.com/episodes/4336`. [Online; accessed 2-December-2012].

Fishman, Y. I. (2009). Can science test supernatural worldviews? *Science and Education*, **18**(6-7), 813–837.

Galak, J., LeBoeuf, R. A., Nelson, L. D., and Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology*, **103**(6), 933 – 948.

Goode, E. (2013). Paranormalism and pseudoscience as deviance. In M. Pigliucci and M. Boudry, editors, *Philosophy of Pseudoscience: Reconsidering the Demarcation Problem*. University of Chicago Press, Chicago. [Retrieved from `http://www.eblib.com`].

Gottipati, S., Qiu, M., Sim, Y., Jiang, J., and Smith, N. A. (2013). Learning topics and positions from debatepedia. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1858–1868.

Hansson, S. O. (2014). Science and pseudo-science. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2014 edition.

Hardisty, E., Boyd-Graber, J., and Resnik, P. (2010). Modeling perspective using adaptor grammars. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 84–292.

Hume, D. (2012). *An Enquiry Concerning Human Understanding*. Start Publishing, New York. [Retrieved from `http://www.eblib.com`].

Hyman, R. and Honorton, C. (1986). A joint communiqu: The psi ganzfeld controversy. *The Journal of Parapsychology*, **50**(4), 351. Last updated - 2013-02-22.

Kelly, T. (2008). Evidence. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2008 edition.

Klebanov, B. B., Beigman, E., and Diermeier, D. (2010). Vocabulary choice as an indicator of perspective. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 253–257.

Kuhn, H. W. (2005). The hungarian method for the assignment problem. *Naval Research Logistics*, **52**, 7–21.

Lewandowsky, S., Gignac1, G. E., and Oberauer, K. (2013). The role of conspiracist ideation and worldviews in predicting rejection of science. *PLOS ONE*, **8**(10), 1–11.

Lin, W.-H., Wilson, T., Wiebe, J., and Hauptmann, A. (2006). Which side are you on? identifying perspectives at the document and sentence levels. In *Proceedings of the 10th Conference on Computational Natural Language Learning*, pages 109–116.

Lin, W.-H., Xing, E., and Hauptmann, A. (2008). A joint topic and perspective model for ideological discourse. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.

Loxton, D. and Prothero, D. R. (2012). *Abominable Science : Origins of the Yeti, Nessie, and other Famous Cryptids*. Columbia University Press, New York. [Retrieved from `http://www.eblib.com`].

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

National Research Council Committee on America's Climate Choices (2011). *America's Climate Choices*. The National Academies Press.

Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 841–848. MIT Press.

Novella, S. (2013). About sbm. `http://www.sciencebasedmedicine.org/about-science-based-medicine/`. [Online; accessed 14-April-2014].

Paul, M. J., Zhai, C., and Girju, R. (2010). Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 66–76.

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., and Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition*, **123**(3), 335 – 346.

Pigliucci, M. and Boudry, M. (2013). Why the demarcation problem matters. In M. Pigliucci and M. Boudry, editors, *Philosophy of Pseudoscience: Reconsidering the Demarcation Problem*. University of Chicago Press, Chicago. [Retrieved from `http://www.eblib.com`].

Quine, W. V. and Ullian, J. S. (1978). *The Web of Belief*. Random House, New York.

Rao, D. and Yarowsky, D. (2010). Detecting latent user properties in social media. In *Proceedings of the NIPS workshop on Machine Learning for Social Networks (MLSC)*.

Sagan, C. (1996). *The demon-haunted world: Science as a candle in the dark*. Ballantine Books, New York.

Shang, A., Huwiler-Mntener, K., Nartey, L., Jni, P., Drig, S., Sterne, J. A. C., Pewsner, D., and Egger, M. (2005). Are the clinical effects of homoeopathy placebo effects? comparative study of placebo-controlled trials of homoeopathy and allopathy. *Lancet*, **366**, 726–732.

Sim, Y., Acree, B. D. L., Gross, J. H., and Smith, N. A. (2013). Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 91–101.

Stark, A., Shafran, I., and Kaye, J. (2012). Hello, who is calling?: Can words reveal the social nature of conversations? In *2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 112–119.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Jounal of Language and Social Psychology*, **29**(1), 24–54.

Thamrongrattanarit, A., Pollock, C., Goldenberg, B., and Fennell, J. (2013). A distant supervision approach for identifying perspectives in unstructured user-generated text. In *International Joint Conference on Natural Language Processing*, pages 922–926.

Wang, S. and Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

Vita

James Spencer Evans is from Falls Church, Virginia. He recieved his B.A. in 2011 from the College of William and Mary. Later that year, he enrolled in the Linguistics graduate program at the University of Texas, where he studied linguistics and natural language processing.

Permanent E-mail Address: jimevans87@gmail.com

This thesis was typeset with LaTeX 2ε[1] by the author.

[1] LaTeX 2ε is an extension of LaTeX. LaTeX is a collection of macros for TeX. TeX is a trademark of the American Mathematical Society. The macros used in formatting this thesis were written by Dinesh Das, Department of Computer Sciences, The University of Texas at Austin, and extended by Bert Kay, James A. Bednar, Ayman El-Khashab, and Nicholas Gaylord. Additional macros used for typesetting this thesis are written by David Walden (`http://walden-family.com/public/texland/ellipses.pdf`), Gonzalo Medina (`http://tex.stackexchange.com/questions/20267/how-to-construct-a-confusion-matrix-in-latex`), and the author.