Spring 2012

Sentiment analysis within and across social media streams

Yelena Aleksandrovna Mejova
University of Iowa

SENTIMENT ANALYSIS WITHIN AND ACROSS SOCIAL MEDIA STREAMS

by

Yelena Aleksandrovna Mejova

An Abstract

Of a thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Computer Science
in the Graduate College of
The University of Iowa

May 2012

Thesis Supervisor: Professor Padmini Srinivasan

ABSTRACT

Social media offers a powerful outlet for peoples thoughts and feelings – it is an enormous ever-growing source of texts ranging from everyday observations to involved discussions. This thesis contributes to the field of *sentiment analysis*, which aims to extract emotions and opinions from text. A basic goal is to classify text as expressing either positive or negative emotion. Sentiment classifiers have been built for social media text such as product reviews, blog posts, and even Twitter messages. With increasing complexity of text sources and topics, it is time to re-examine the standard sentiment extraction approaches, and possibly to re-define and enrich the definition of *sentiment*. Thus, this thesis begins by introducing a rich, multi-dimensional model based on Affect Control Theory, which shows its usefulness in sentiment classification. Next, unlike sentiment analysis research to date, we examine sentiment expression and polarity classification within and across various social media streams by building topical datasets within each stream. When comparing Twitter, reviews, and blogs on consumer product topics, we show that it is possible, and sometimes even beneficial, to train sentiment classifiers on text sources, which are different from the target text. This is not the case, however, when we compare political discussion in YouTube comments to Twitter posts, demonstrating the difficulty of political sentiment classification. We further show that neither discussion volume nor sentiment expressed in these streams correspond well to national polls, putting in question recent research linking the two. The complexity of political discussion also calls for a more specific re-definition of “sentiment” as agreement with the author’s political stance. We conclude that sentiment must be defined, and tools for its analysis designed, within the larger framework of human interaction.

Abstract Approved: ________________________________
Thesis Supervisor

____
Title and Department

____
Date

SENTIMENT ANALYSIS WITHIN AND ACROSS SOCIAL MEDIA STREAMS

by

Yelena Aleksandrovna Mejova

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Computer Science
in the Graduate College of
The University of Iowa

May 2012

Thesis Supervisor: Professor Padmini Srinivasan

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Yelena Aleksandrovna Mejova

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Computer Science at the May 2012 graduation.

Thesis Committee:

Padmini Srinivasan, Thesis Supervisor

Alberto Maria Segre

Juan Pablo Hourcade

G. R. Boynton

Alison Bianchi

ABSTRACT

Social media offers a powerful outlet for peoples thoughts and feelings – it is an enormous ever-growing source of texts ranging from everyday observations to involved discussions. This thesis contributes to the field of *sentiment analysis*, which aims to extract emotions and opinions from text. A basic goal is to classify text as expressing either positive or negative emotion. Sentiment classifiers have been built for social media text such as product reviews, blog posts, and even Twitter messages. With increasing complexity of text sources and topics, it is time to re-examine the standard sentiment extraction approaches, and possibly to re-define and enrich the definition of *sentiment*. Thus, this thesis begins by introducing a rich, multi-dimensional model based on Affect Control Theory, which shows its usefulness in sentiment classification. Next, unlike sentiment analysis research to date, we examine sentiment expression and polarity classification within and across various social media streams by building topical datasets within each stream. When comparing Twitter, reviews, and blogs on consumer product topics, we show that it is possible, and sometimes even beneficial, to train sentiment classifiers on text sources, which are different from the target text. This is not the case, however, when we compare political discussion in YouTube comments to Twitter posts, demonstrating the difficulty of political sentiment classification. We further show that neither discussion volume nor sentiment expressed in these streams correspond well to national polls, putting in question recent research linking the two. The complexity of political discussion also calls for a more specific re-definition of “sentiment” as agreement with the author’s political stance. We conclude that sentiment must be defined, and tools for its analysis designed, within the larger framework of human interaction.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
INTRODUCTION

Over the years, surveys have been the main method for answering the question *what do people think?* A careful sampling of the polled population and a standardized questionnaire have become the standard way of learning about large groups of people [71]. Recently though, the era of wide-spread internet access and social media has brought a new way of learning about large populations. This thesis contributes to a field of **Sentiment Analysis** (SA), which aims to extract emotions and opinions from text, and most notably from social media. Given, for example, Twitter messages about a local event, or blog posts about an issue, or reviews of the latest camera, the goal of SA is to classify the emotions expressed in these texts along a polarity spectrum of positive - neutral - negative. A more advanced classification task would be to consider multiple emotional states like “disappointed”, “excited”, or “angry”.

In the past decade, sentiment analysis has become a hot research field and a booming industry. For instance, IBM SPSS[1] provides quantitative sentiment summaries of survey data to assist businesses in understanding consumer attitudes. LexisNexis[2] compiles consumer confidence and brand perception summaries using news media, while OpSec[3] also mines user-generated data (social media). Wall Street has

[1] http://www-01.ibm.com/software/analytics/spss/

[2] http://www.lexisnexis.com/risk/data-analytics.aspx

[3] http://opsecsecurity.com/brand-protection/online-brand-protection/sentiment-analysis

also started to use SA in their trading algorithms with companies like OpFine[4] providing up-to-date sentiment tracking of financial news. Even several major news sources like The Washington Post[5] and Politico[6] now provide social media statistics on popular political figures.

Much of early SA research centered around product reviews, such as ones left for products on Amazon.com, defining sentiment as either positive, negative, or neutral. These were a convenient source of labeled data, as star ratings were used as quantitative indicators of the author's opinion. Later, annotated datasets were created for more general types of writing such as blogs, web pages and news articles. Recent growth of Twitter has produced a plethora of research tracking topics and sentiment for all kinds of new applications: [1], for example, try to predict box-office revenues of movies, [10] track H1N1 epidemics, and [73] monitor effects of an earthquake, all using Twitter.

Although exciting in their diverse applications, most sentiment analysis studies have focused on one social media source, tailoring their approaches to a subset of a wide variety of texts. Furthermore, analysis of political discussions proves to be quite challenging, putting in question whether conceptualization of sentiment and lexicon-based approaches developed for mining product reviews are suitable for analysis of such rich discourse.

[4] http://www.opfine.com/

[5] http://www.washingtonpost.com/politics/mention-machine

[6] http://news.cnet.com/8301-13772_3-57358111-52/politico-to-mine-facebook-for-insight-into-voter-sentiments/

Motivated by the above observations, this thesis addresses the following four questions:

1. ***Is it possible to enrich the definition of sentiment?***

In Chapter 3, we introduce a rich sentiment model based on a theory from Sociology – Affect Control Theory – that postulates that affective meaning can be expressed as a point in a multi-dimensional semantic space and provides empirically-derived equations for understanding the affective meanings of words according to the context in which they appear. We show that it has the potential to expand SA's current simplistic view of sentiment and improve polarity classification performance.

When examining political texts in Chapter 6 we also show the distinction between positive/negative emotion and agreement with a political stance, both of which can be considered as a kind of "sentiment".

2. ***Which document representation approaches are the best for building data-driven sentiment classifiers?***

We develop a set of guidelines for document representation used for sentiment polarity classification. We conduct an experiment involving three popular (de facto standard) SA datasets and compare popular feature definition and selection techniques, which include standard IR techniques like stemming and feature weighting schemes, advanced NLP techniques like phrase chunking and n-gram parsing, and task-specific ones like negation enrichment. A variety of performance metrics and cost analysis of memory and running time highlight the merits of these approaches and the costs of using them.

3. ***What are the differences and similarities between the expression of sentiment in different social media streams?***

We examine different social media sources for the purposes of sentiment analysis in two chapters: one comparing reviews, blogs, and Twitter documents on a set of common consumer product topics (Chapter 5) and comparing YouTube comments and Twitter posts on a set of political topics (Chapter 6). In both studies we create annotated datasets which reveal stream-specific topical sentiment peculiarities, and perform set of classification experiments evaluating the extent to which information learned from one source is useful in classifying another.

4. ***What is political discourse in social media like, and is it indicative of national political sentiment?***

We explore the task of analyzing political speech in social media, first in YouTube and Twitter (Chapter 6) and then in Twitter alone (Chapter 7). Annotating for both sentiment and stylistic features, such as humor, sarcasm, and quoting, we examine the nature of political expression for both few vocal power-users and the more reserved majority. Using political sentiment classifiers, we track sentiment change around political debates and compare it to the national polls, uncovering biases social media users may have.

Addressing both the fundamental questions of sentiment analysis and pushing the frontiers of political sentiment extraction, this thesis provides insights into methodological approaches to extracting emotion from text as well as the nature of sentiment expressed in various social media streams, and especially in political domain. With the growing popularity of social media, and interest from both businesses and media, sentiment analysis of user-generated data is not only an interesting case of text analysis, but a research area with bright and interesting future.

CHAPTER 2
RELATED WORK

When conducting serious research or making every-day decisions, we often look for other people's opinions. We consult political discussion forums when casting a political vote, read consumer reports when buying appliances, ask friends to recommend a restaurant for the evening. And now the Internet has made it possible to find out the opinions of millions of people on everything from latest gadgets to political philosophies. Social media now commands over 22% of the world's total time spent online[1] with 65% of adult internet users using some kind of social networking site[2]. The Internet is increasingly both the forum for discussion and source of information for a growing number of people.

As a response to the growing availability of informal, opinionated texts like blog posts and product review websites, a field of Sentiment Analysis has sprung up in the past decade to address the question *What do people* ***feel*** *about a certain topic?* Bringing together researchers in computer science, computational linguistics, data mining, psychology, and even sociology, sentiment analysis expands the traditional fact-based text analysis to enable opinion-oriented information systems.

[1] http://blog.nielsen.com/nielsenwire/global/social-media-accounts-for-22-percent-of-time-online/

[2] http://pewinternet.org/Reports/2011/Social-Networking-Sites.aspx

2.1 What is Sentiment?

One of the challenges of Sentiment Analysis is defining the objects of the study – opinions and subjectivity. Originally, subjectivity was defined by linguists, most prominently, Randolph Quirk [66]. Quirk defines *private state* as something that is not open to objective observation or verification. These private states include emotions, opinions, and speculations, among others. Wiebe, a prominent Natural Language Processing (NLP) researcher, used Quirk's definition of the *private state* when tracking point of view in narrative [88]. She defines private state as a tuple `(p, experiencer, attitude, object)` relating experiencer's state p to his/her attitude possibly toward an object. In practice, a simplified version of this model, where we look only at polarity and the target of the sentiment, is usually used. In fact, many researchers define sentiment loosely, as a negative or positive opinion [62, 30, 52]. Some researchers use products that provide pre-compiled lists of words in various groupings, some of which are related to emotional states. These include Linguistic Inquiry and Word Count (LIWC)[3] and Profile of Mood States (POMS)[4]. In Chapter 3, we introduce a rich, empirically-derived sentiment model based on a theory from Sociology – Affect Control Theory – and show that it has the potential to expand SA's current simplistic view of sentiment.

[3] http://www.liwc.net/

[4] http://www.mhs.com/product.aspx?gr=cli&id=overview&prod=poms

2.2 Polarity Classification

A basic and typical task in sentiment analysis is *text polarity classification*, where the classes of interest are *positive* and *negative*, sometimes with a middle *mixed* class. There are two approaches to this problem: one may use a sentiment lexicon (a list of words with known sentiment polarity) as in LIWC or POMS as described above, or one may build a “model” of the language used for each polarity using training data. Although simple and easy to use, the lexicon-driven approach is inflexible in the light of the diversity of topics and styles of writing – the lexicon words may simply not appear in the text of interest or may be used in a peculiar fashion. However, using machine learning techniques, one can build a classifier that is specifically trained on a particular text, and thus captures the peculiarities of the language used in it.

A variety of ways have been used to represent text for this purpose. The most common is a bag-of-words representation whereby each word becomes a feature, having binary value (1 if it appears in document, 0 if it does not) or some other value (such as the number of times it appears in the document) [60]. More complex representations include n-grams (n number of consecutive words) [12], phrases (identified using parts of speech), and negation-enriched words (differentiating *“bad”* from *“not bad”*) [13]. These and many other techniques for representing the documents (defining the *“feature space”*) have been proposed in the literature.

However, because of a lack of standard datasets and approaches, some studies have produced conflicting results, or ones which were not directly comparable. One may consider whether a unigram representation is sufficient, or whether further computation should be done to generate 2- or 3-grams to represent the text. Upon consulting SA literature one discovers that [62] find that bigrams do not improve performance beyond that of unigram presence. Yet, in a study by [12] of a much larger corpus, higher order n-grams do show marginal improvement in performance. A third study by [15] shows performance increase for 2- and 3-grams, while it is unclear which is better. Similarly, negation-enriched features have been used by [15] and [62], but with conflicting results. Term weighting has been examined in small datasets by [62] and [60]. The latter find that term occurrence outperforms un-normalized term frequency, but not normalized term frequency. In Chapter 4 we test the common feature definition and selection techniques on a set of standard datasets of varying sizes and produce guidelines for building a general-purpose sentiment classifier.

2.3 Mining Social Streams

One of the reasons sentiment analysis has become so prominent in the last decade is the rise of social media. Product reviews have been a common source of data for SA, since the star rating provided a quantitative label for the documents (making it unnecessary to manually label them) [62]. These were followed by annotated datasets of blogs [52], web pages [36] and news articles [83], followed by a flurry of Twitter-related research ranging from predicting box-office revenues of movies [1], to tracking H1N1 epidemics [10], to monitoring effects of an earthquake [73].

Few studies have compared sentiment expression and classification in different social media sources. [4] examine classification performance on review and blog sources in which documents are constrained in length, finding that it is easier to classify shorter documents than their longer counterparts. And [64] develop iterative algorithms for filtering noisy data during source adaptation from Twitter and Blippr (a micro-review site) to movie reviews. More generally, [14] explores the use of a variety of sources, including blogs, news, usenet, and conversational telephone speech for sequence labeling tasks, such as named-entity recognition, shallow parsing, and part-of-speech recognition. Cross-language sentiment classification has also been performed by [86] on Chinese and English reviews. However, no cross-stream comparison has been done while controlling for the topical coverage of the datasets. In Chapters 5 and 6 we build two annotated multi-source datasets and compare sentiment they express and classification models they generate.

2.4 Political Sentiment

The recent role of social media in political actions in US, Middle East, and elsewhere around the world has produced a gamut of studies on mining of political speech online. A report on Social Media in the Arab World recognizes "the pivotal role of the microblogging [Twitter] site [...]" and "the role that social media will continue to play in Tunisia, Egypt, and the rest of the Arab world" [25]. Thus, from tracking discussions of political debates [19] to predicting election outcomes [84], social media has become a gold mine for political sentiment research. For example, [44] use social media to determine whether news sources are biased in favor of covering one political party more than another. Focusing on representation of political figures in Twitter,

[67] have developed a way to detect *astroturf* (politically-motivated speech which creates appearance of widespread support for a candidate or opinion). Elections have been studied through the lens of social media: [47] examine the usage patterns of social media by US political parties in the 2010 Midterm Election, whereas [22] look at the conversations surrounding German political parties during the 2009 Federal Elections.

However, some researchers are skeptical whether it is really possible to automatically analyze political speech [24], questioning the efficacy of current techniques to tackle one of the most diverse and convoluted writing styles. Metaxes et al. [53], for example, find that electoral predictions using various previously published methods on Twitter data is no better than chance. Among these techniques are discussion volume, lexicon-driven sentiment classification, and user-specific political leaning estimation. Chapter 7 examines political speech on Twitter, focusing on the Republican candidates for the 2012 US Presidential nomination, in attempt to better understand the kind of political writing social media contains, and how well it corresponds to the national political polls.

* * *

Set in this research context, we explore the questions raised in our Introduction in the following chapters.

CHAPTER 3
REDEFINING SENTIMENT

We begin our exploration of Sentiment Analysis by examining the very definition of sentiment. Throughout SA research, sentiment has been defined as positive or negative (sometimes neutral or mixed classifications are added). In this chapter, we ground the semantic modeling of emotion in Affect Control Theory – a sociological theory which provides multi-dimensional view of affective emotion, as well as means to combine meanings of individual words in a sentence to produce a higher-level emotional summary of the described event. We show that polarity classifiers that use ACT lexicons outperform those that use standard SA ones. Moreover, we show that the ACT equations that modify the affect scores of words according to their context significantly improve performance. Finally, we propose several avenues of future research in hopes that the two fields that so far have been quite separate can benefit each other.

3.1 Affect Control Theory

For half of a century, human emotion has been quantitatively studied by sociologists. The meaning of words has been central in the Symbolic Interactionism paradigm, which states that people act toward other people and things based on the meanings that they have given to them [27]. Language plays an important part of this interaction as a means of negotiating meaning through symbols. The meaning of words includes the emotional responses they evoke.

The study of the affective meanings of words has been a central part of Affect Control Theory (ACT) research. This theory postulates that there are certain cultural norms that dictate the affective meanings of words, norms that people in a culture with a common language share [70]. Even out of context, each word has an affective meaning (called a *fundamental*), but this meaning changes when put in context (becoming a *transient*). When encountering a situation, people gather the fundamental meanings, then adjust these in the light of their context, and then act accordingly.

In order to quantify these meanings, Osgood, Suci and Tannenbaum [59] use a form of the *semantic differential* technique, which pinpoints meanings in a multidimensional *semantic space*. Each dimension is a scale like *fast* versus *slow*, *good* versus *bad*, *hard* versus *soft*. Using factor analyses, they determined that three dimensions were the most important in differentiating meanings, reducing the semantic space to a three-dimensional cube. These dimensions are *Evaluation* (*good* vs. *bad*), *Potency* (*strong* vs. *weak*), and *Activity* (*lively* vs. *inactive*). (Notice that ACT Evaluation dimension is the customary SA polarity.) Using these three dimensions, many studies have been performed to measure the affective meanings various people associate with words. Culture-specific lexicons have been compiled for several countries, including United States, Canada, Japan, Germany, China, and Northern Ireland [70] and subcultures, including Internet users [38], state troopers [29], and religious groups [76].

This multi-dimensional quantification of affect now allows us to relate the affective meanings of various words to each other. Using the data gathered on thousands of sentences, ACT researchers have extrapolated systems of linear equations which combine the affective meanings of individual words (*fundamentals*) to produce new context-dependent meanings for the same words (*transients*). For example, below are the equations for recalculating the *evaluation* score for concepts in a sentence of form Actor-Behavior-Object by Heise (1969) [28]:

$$A'_e = -0.15 + 0.37A_e + 0.55B_e + 0.07O_e + 0.25B_eO_e$$

$$B'_e = -0.24 + 0.23A_e + 0.60B_e + 0.07O_e + 0.25B_eO_e$$

$$O'_e = -0.13 + 0.17A_e + 0.40B_e + 0.36O_e + 0.30B_eO_e$$

where A'_e, B'_e, and O'_e are the new *evaluation* scores for the Actor, Behavior and Object respectively. Unlike simple combination (that is, averaging) of the sentiment contained in text which is used in Sentiment Analysis, these equations can also take into consideration the other three dimensions. Below is the equation for the *evaluation* dimension in a sentence of form Actor-Behavior-Object for the Actor from a different study [75]:

$$\begin{aligned}
A'_e \quad = \quad & -0.98 + 0.468A_e - 0.015A_p - 0.015A_a + 0.425B_e \\
& -0.069B_p - 0.106B_a + 0.055O_e - 0.0205O_p - 0.0015O_a \\
& +0.048A_eB_e + 0.130B_eO_e + 0.027A_pB_p + 0.068B_pO_p \\
& +0.007A_aB_a - 0.038A_eB_p - 0.010A_eB_a + 0.013A_pB_e \\
& -0.014A_pO_a - 0.058B_eO_p - 0.070B_pO_e - 0.002B_pO_a \\
& +0.010B_aO_e + .019B_aO_p + 0.026A_eB_eO_e \\
& -0.006A_pB_pO_p + 0.031A_aB_aO_a \\
& +0.033A_eB_pO_p + 0.018A_pB_pO_p
\end{aligned}$$

This equation expresses various interactions between the affective meanings of all three concepts of the event. For example, the large positive coefficient of the B_eO_e term captures the idea that actors seem especially nice if they behave nicely towards good others (or badly towards nasty others). On the other hand, the negative B_eO_p coefficient captures the idea that actors seem nicer when they treat others that are weak nicely or are less positive toward strong others. This can be thought of as a *social responsibility* norm [70]. The most striking fact is that these complex interactions are inferred automatically from the collected data, not hand crafted.

Finally, it is possible to model the actor's reaction to a particular situation. In order to do this, the theory defines another measure, *deflection*, which is the Euclidean distance between the fundamental cultural sentiments and the transient impressions [70] (how much the meaning of words changes in a given context). For example, a sentence *"Mother beats the child"* would have a large deflection, since it does not correspond to the societal role we attribute to mothers (generally considered to be caring and gentle to the child). Notice that two out of three words in the sentence – *"mother"* and *"child"* – are relatively positive, whereas *"beat"* is negative. So, a simple averaging of the *evaluation* dimensions would produce an overall positive score. But if we re-evaluate the affective meaning of each of the words in their context first, the scores of individual words will reflect the meaning of the sentence, and the summation of the new *evaluation* scores would reflect the negativity of the sentence.

3.2 Example

To illustrate the difference between the standard Sentiment Analysis approach and a Affect Control Theory approach, we calculate the polarity score for the sentence below:

The nurse poisoned her nephew.

Table 3.1: ACT affective dimensions for select words and SWN positive/negative entries

word	$Evaluation$	$Potency$	$Activity$	SWN_p	SWN_n
nurse	2.05	1.01	0.84	0.38	0.13
poison	-3.02	0.71	-0.44	0.00	0.25
nephew	1.20	-0.37	1.45	0.13	0.00

Table 3.1 shows the entries for these words in our lexicons. The first three columns are the *evaluation*, *potency* and *activity* scores from the ACT lexicon and the last two are positive and negative scores from the SentiWordNet lexicon. The scale of ACT scores is $[-3, 3]$, and the scale of the SentiWordNet $[-1, 1]$ (see next section for more information on these lexicons). Notice that the ACT scores implicitly contain the polarity information, whereas SentiWordNet scores have two separate values for each polarity. That is, *nurse* has both positive (0.38) and negative (0.13) scores, though more positive than negative.

To calculate the overall polarity score of the sentence, we can use the standard Sentiment Analysis approach of summing over the polarity scores of all words in it. Using ACT *evaluation* scores we get 0.23, a mildly positive score. To use SentiWordNet scores, we first need to get a single score for each word by subtracting the negative from the positive. This way the sentence is evaluated at 0.13, also a mildly positive score. Because two out of three of the words are positive, a simple addition of the scores misses the point that someone (*nurse*) has done something awful (*poison*) to somebody (*nephew*).

But now we can first adjust the polarity scores of the words using ACT equations. Using Heise study from [28], we get the new *evaluation* scores for *nurse*: -1.87, *poison*: -2.41, and *nephew*: -1.64. Summed together, we get a sentence polarity score of -5.92, one that strongly reflects the negativity of the sentence. Furthermore, if we use more elaborate equations from Smith-Lovin [75], we get the following scores: *nurse* (-1.63), *poison* (-2.35), and *nephew* (0.06), resulting in an overall sentence score of -3.92. Using either equation, because of the re-definition of each word within its context, the summary of the polarities of the constituent words in the sentence reflects its meaning more accurately.

3.3 Experiments

In order to compare sentence polarity summaries produced using a standard Sentiment Analysis lexicon to those produced using Affect Control Theory lexicons, we perform a set of experiments. We use an automatically annotated extension of WordNet – SentiWordNet [21] – as a standard Sentiment Analysis lexicon. It has been used to classify financial news [18] and news headlines [8]. The Affect Control Theory lexicon was compiled using the INTERACT[1] system, and included lexicons collected in eight studies conducted in the span between 1977 and 2003. Because several studies used the same vocabulary, the EPA scores for words which were used in several studies were averaged. The lexicons also were divided into two genders (female and male), and the scores were averaged over the two versions. The final lexicon consisted of 1886 Identities (nouns and noun phrases) and 1009 Behaviors (verbs and verb phrases).

In order to examine the kinds of polarity scores standard SA and ACT lexicons provide, we selected 150 most negative, 150 most positive, and 100 neutral words from both Identity and Behavior sets (that is, 400 of each). Using these, $10,000$ word triples of form Actor-Behavior-Object were created by random sampling of

[1] http://www.indiana.edu/˜socpsy/ACT/interact/

the lexicon. Three scores were attained from the ACT lexicon: one which sums the original *evaluation* scores (ACT_{simple}), one which sums the scores modified using Heise (1969) [28] (ACT_{Heise}), and one which sums the scores modified using Smith-Lovin (1987) [75] ($ACT_{Smith-Lovin}$). To determine the final polarity we consider negative scores as an indication of negative polarity and non-negative scores as indication of positive polarity.

The correlation between the two modified scores ACT_{Heise} and $ACT_{Smith-Lovin}$ was 0.883, with 89.36% of the scores matching. The original score correlated less with ACT_{Heise} (0.661), and more with $ACT_{Smith-Lovin}$ (0.863). There is almost no correlation between the SWN scores and the ACT scores. Out of the triplets for which terms were found in the SWN lexicon (5016), only around 56% of the SWN scores matched with ACT.

To check the quality of the ratings, 800 triples were randomly selected for manual rating. The triples were classified by two annotators, and a third annotator broke the ties. Out of the 800 triples 94 were judged non-sensical (not surprising, considering the triples were synthetic). Below are some examples of the ones that did make sense:

bridegroom kisses a crony

robber double crosses an old maid

junior college student rebels against a preacher

The inter-annotator agreement between the two main annotators was measured using Cohen's Kappa [40]: $\kappa = 0.699$ signifying a substantial amount of agree-

Table 3.2: Performance of polarity classifiers using various lexicons on 716 annotated triples

Lexicon	**Acc**	P_{pos}	P_{neg}
SWN	0.548	0.444	0.635
ACT_{simple}	0.719	0.642	0.775
ACT_{Heise}	0.782	0.765	0.791
$ACT_{Smith-Lovin}$	0.803	0.857	0.781

ment. Because the triples are synthetic (that is, they were produced programmatically), this is surprisingly high.

Table 3.2 shows the performance of the polarity classifiers using a standard SA lexicon (SWN) and three versions of the ACT lexicon. The first measure is Accuracy, the two others are precision for the positive class (P_{pos}) and negative class (P_{neg}). All ACT-driven classifiers outperform SWN-based one by a large margin. We see further improvement when the context-incorporating equations are applied (at significance level of $p < 0.01$), though there is little distinction between the two equations ($p = 0.1635$).

In many sentences the polarity of the Behavior (the verb) did not coincide with that assigned to the whole triple by the annotators. In these instances, ACT-driven classifier labeled 43 instances correctly, which SWN-driven one mis-classified. For example, the sentence *"junior college student honors a skirt chaser"* contains a positive Behavior ($B_e = 1.94$), but a negative Object ($O_e = -2.34$). Because doing something good to a bad object is usually considered bad, when we apply Heise equation to *"honors"*, the evaluation score changes to $B_e = -0.40$, favoring an overall negative evaluation of the phrase. SWN-driven classifier, on the other hand produced an overall positive evaluation of 0.63.

3.4 Conclusions

3.4.1 Summary of Findings

The sophistication and empirical nature of Affect Control Theory lexicons and analysis tools surpass the much simpler definition and treatment of affective meaning in Sentiment Analysis. In this chapter we incorporate Affect Control Theory resources into the Sentiment Analysis task of polarity classification, and show that they produce more accurate polarity judgments. Although the task considered here is a standard SA one, similar techniques can be used to extract the other two ACT dimensions (potency and activity).

However, the ACT resources available to date are still limited. The near-3,000 word lexicon we created using the ACT studies is still limited compared to vocabularies of social media datasets which may span in millions of words. Thus, in the rest of this thesis we use data-driven approaches, as described in Chapter 4.

3.4.2 Future Work

Future research in incorporating these two areas are promising to be fruitful. First, the culture-specific lexicons produced by various ACT studies provide high quality multi-dimensional annotations for the automatic text analysis. Furthermore, by leveraging unsupervised lexicon building techniques [32, 49] these lexicons may be extended, to benefit both sociologists and text mining researchers. Second, ACT equations allow us to evaluate concepts in their textual context, introducing a more involved semantic processing of text. Finally, we may now apply these techniques to a wide variety of socially-generated text available on the web, expanding the scope of a typical sociological study from hundreds or thousands to millions of subjects.

CHAPTER 4
DATA REPRESENTATION

As described in the previous chapter, text polarity classification is one of the main tasks of Sentiment Analysis, but unlike the lexicon-driven approach we adopted there, this task can instead be approached by building a “model” of sentiment polarities using some training data. These training data must first be processed and expressed as a set of “features”. Much has been written on the usefulness of various feature definition techniques for this task. However, it is still unclear which features are the best. To better understand the merit of current techniques, we study features for sentiment analysis along three dimensions. First, we examine the basic units extracted from texts: words, n-grams, and phrases. Second, we explore feature selection, considering both frequency-based and probabilistic strategies. Third, we explore feature generalization. Here, besides parts of speech (POS) we explore three different lexicons: one extracted from Affect Control Theoretical [58] sociological studies of emotion [50], and two extensions of WordNet: SentiWordNet [21] and WordNet-Affect [80]. This third portion of our study focuses on testing specific hypotheses that underly many of the feature definition methods observed in SA research.

We test these techniques on three datasets: an IMDB movie review set from [61], a new product review dataset that has previously been used for review spam detection [34], and a multi-domain dataset from [6]. The first dataset is standard, although rather small (2,000 documents). The two others are somewhat more realistic (tens of thousands of documents). We show that the size of the dataset affects the performance of some of the techniques. For example, using top few thousand features using Mutual Information for large datasets improves the performance, whereas it proves too selective for the smaller dataset.

Finally, because a marginal improvement in performance may be overshadowed by the cost of computing the feature, we present cost analysis for each of the features in terms of processing time and storage space.

4.1 Related Work

As a systematic study of various popular features for sentiment polarity classification, we briefly discuss each here, but for a comprehensive discussion of these topics, see [63].

A main concern in automatic document classification is representation. The standard "bag of words" representation, consisting of vectors of the terms the document contains, has been used as a baseline in [62] and is still widely used today. Term weighting strategies specifically for polarity classification have been thoroughly examined by [60], who tested classifier performance on variations of the classic TFIDF scheme. They confirm results from [62], showing that binary features (representing word occurrence) outperform features weighted by raw term frequency. We check the consistency of these findings on the dataset these studies use and also explore their robustness with two larger datasets.

N-grams capture some of the context around individual words. However, it is still unclear whether n-grams are useful in polarity classification. [62] find that bigrams do not provide an improvement in performance over a unigram (bag of words) baseline. On the other hand, [12] experiment on a much larger dataset (over 320k product reviews), and show significant improvements with higher-order n (up to 6). Though statistically significant, overall precision, recall, and F_1 measure improvements are very slight ($< 2\%$). We address this question, examining whether the increased computation and feature space n-grams require provide an improved performance.

Lexicons, both constructed manually or automatically, have been used to determine the polarity of text. Manually-constructed lexicons include one created by [81], in which words are associated with affect categories, specifying intensity (strength of affect level) and certainty (degree of relatedness to the category). For example, the word "*contempt*" may be annotated with the class *repulsion* with intensity of 0.7 and centrality of 0.6 (both on a scale from 0.0 to 1.0). [21] and [80] use a lexical database WordNet to build sentiment-annotated resources: SentiWordNet and WordNet-Affect, respectively. SentiWordNet has been used to classify financial news [18] and news headlines [8], and WordNet-Affect to classify music lyrics [31] as well as congressional floor debates [2]. Both of these lexicons are used in our research. We also test the Affect Control Theory-based lexicon developed in the previous chapter.

Part-of-speech information has been successfully used by SA researchers. [54] use *semantic orientation* of adjectives, i.e. a measure of the positive or negative sentiment expressed by a word. [87] use four adjectival appraisal groups: Attitude, Orientation, Graduation, and Polarity. Other parts of speech have been found to be useful, such as adverbs [3], nouns [56], and verbs [89]. An especially useful part of speech for the task of polarity classification is negation. Because it reverses the polarity of the word it is applied to, it is beneficial to take it into consideration when using a polarity lexicon. To supplement the standard bag-of-words document representation, [13] create special features representing negated words (for example, *like-NOT*). We explore POS in general as well as the categories of adjectives, verbs, nouns as also negation.

Attempts have been made to create more semantically cohesive features than n-grams. [68] use a subsumption hierarchy to identify n-grams and extraction pattern features that are strongly associated with opinionated text. These extraction patterns are automatically generated from text using a tool called AutoSlog [69]. Resulting patterns are generalized phrases such as "*drive <NP> up the wall*", which expresses the sentiment of annoyance, even though the words in the phrase "drive", "up" and "wall" are not themselves opinionated. Such patterns have also been extracted using bootstrapping. [49] propose a Weighted Mutual Exclusion Bootstrapping (WMEB) algorithm for extracting semantic lexicons and templates for multiple categories. In our study, we use a part-of-speech tagger to extract phrases from text as an alternative to n-grams. Importantly, we use parts-of-speech and lexicons to generalize these features, similarly to generalizations of dependency relation triples in [35]. These generalizations tests allow us to study implicit assumptions made in the literature concerning both parts-of-speech and lexicons.

4.2 Features

In this section, we present the different features and their potential usefulness in polarity classification.

4.2.1 Feature Definition

4.2.1.1 Words and Stems

Though one may certainly represent a document by the raw words in it, a classic technique in information retrieval is to stem the words to their morphological roots. Stemmed feature vectors are smaller in size, they aggregate across occurrences of variants of a given word. Stemming has had mixed success in both information retrieval and text mining. [15], for example, show that stemming produces mixed results on different datasets. They conclude that "corpus of reviews is highly sensitive to minor details of language, and these may be glossed over by the stemmer". An example they observe is that negative reviews tend to occur more frequently in the past tense, since the product might have been returned.

4.2.1.2 Binary versus Term Frequency Weights

A standard approach in information retrieval is to use term frequency (TF) weights to indicate the relative importance of features in document representations. However, some research has shown that binary weighting (0 if the word appears in the document, 1 otherwise) is more beneficial for polarity classification [62]. In a study of the standard information retrieval weighting schemes in SA, [60] found that using binary features is better than raw term frequency, though a scaled TF version performs as well as binary. Thus, we include runs in our experiments which compare the two weighting schemes.

4.2.1.3 Negations

Negations such as *not* and *never* are often included in stopword lists, and hence are removed from the text analysis. Combined with other words, though, negations reverse the polarity of words. Because polarity classification may be affected by negations, SA researchers have tried incorporating them into the feature vector. We take the approach of [13] who use a heuristic to identify negated words and create a new feature by appending *NOT-* to the words (for example, a phrase “don’t like” results in feature *NOT-like*).

4.2.1.4 N-grams

Negation phrases discussed above can be considered as a special case of n-grams, which are ordered sets of words. The benefit of using n-grams instead of single words as features comes in being able to capture some dependencies between the words and the importance of individual phrases. In a study of subjective text fragments, [12] found a significant improvement in polarity classification task using high n (up to 6). However, it is unclear if n-grams are as useful in a smaller dataset where there may not be enough data to capture information about their occurrence patterns. In our experiments, we generate features of up to 3-grams using CMU Toolkit (http://www.speech.cs.cmu.edu).

4.2.1.5 Phrases

Since n-grams are often synthetic, in that they do not necessarily represent a semantically cohesive part of text, we explore the use of grammatical phrases as features. Using a CRF-based phrase chunker (http://jtextpro.sourceforge.net/), we break the text into phrases and use these as features. We further explore phrase features with modifications below.

4.2.2 Feature Selection

4.2.2.1 Frequency-Based Selection

In text modeling, it is often the practice to remove words which appear rarely in the corpus. These are presumed to be perhaps misspellings, that do not help in generalization during classification. On the other hand, words that occur only once in a given corpus have been found to be high-precision indicators of subjectivity [89]. Rare terms, thus, may serve an important role in classification, and so we test various cutoffs using frequency counts.

4.2.2.2 Mutual Information Based Selection

The performance of the classifier may also be improved by removing some of the less useful features. One of the common feature selection measurements is expected Mutual Information [46]. For a binary random variable U and an also binary class variable C it is calculated as follows:

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U = e_t, C = e_c) log_2 \frac{P(U = e_t, C = e_c)}{P(U = e_t)P(C = e_c)}$$

where $e_t = 1$ when the document contains the term t, 0 otherwise, and similarly $e_c = 1$ when the document belongs to the class c, 0 otherwise [48]. Usually the features are scored by the expected MI and top several are taken as the most useful in classification. This is also the approach we take.

4.2.2.3 Part of Speech-Based Selection

In particular for SA, certain POS have been determined to be more useful in classification tasks. For example, [3] show that using adjectives and adverbs works better than using adjectives alone. [9] also use verbs for sentiment classification. If indeed adjectives are important factors in predicting sentiment polarity [63], limiting the feature space to only these may improve classifier performance by removing less useful words. We test this notion by retaining only words that are adjectives, verbs, and nouns individually and in combination.

4.2.2.4 Lexicon-Based Selection

Similarly, sentiment-annotated lexicons may be used for feature selection. By selecting terms which are indicative of strong sentiment, less useful features may be excluded from the feature set. Popular lexicons are the extensions of WordNet (http://wordnet.princeton.edu/), a large lexical database of English. SentiWordNet, for example, contains polarity and objectivity labels for the WordNet terms [21]. In WordNet-Affect [80] take advantage of synsets - word groupings in WordNet - to label each synset with affective labels. Both have been widely used in the community, and we use both lexicons in our analysis. Furthermore, we use Affect Control Theory lexicon from Chapter 3.

4.2.3 Feature Generalization

4.2.3.1 Phrase Generalization (POS-driven)

To avoid the problem of data sparsity we generalize the phrases described earlier by replacing some of the words in each phrase with their POS. The most drastic generalization is replacing all words with POS, though this may remove too much information from the phrase. Instead, we may want to retain words belonging to important POS and generalize others. As discussed above, adjectives, verbs, and nouns may be indicative of sentiment polarity. We explore just how much these POS individually and in combination help in classification by generalizing all words by their POS except for adjectives. Likewise we study verbs and nouns.

4.2.3.2 Phrase Generalization (Lexicon-driven)

We may also wish to generalize phrases by considering sentiment-annotated lexicon words as important. We experiment with the three above lexicons: Affect Control Theory (ACT), SentiWordNet (SWN) and WordNet-Affect (WNA).

4.3 Experimental Setup

We perform tests on three datasets which are described in Table 4.1. First comes from [61] and includes 1000 positive and 1000 negative movie reviews from

IMDB. Second dataset comes from [34] and is a sample of 20,000 product reviews (taken out of 5,838,855 original documents for tractability). We sampled according to the polarity proportions in the original dataset, taking reviews with rating 5 to be positive and 1 to be negative. The third dataset is a subset of another multi-domain sentiment dataset which has been used in [6, 20]. Note that the last two datasets have unequal number of positive and negative reviews.

Classification was done using Weka sequential minimal optimization (SMO) algorithm for training Support Vector Machines (SVM) [65]. We use an SVM for classification for two reasons. First, it is not our intention to determine the best classifier for the task, but the best feature set. Second, SVMs have been widely used in SA and in many cases outperform all other classifiers [41, 62]. Our classifier was tested using 10-fold cross-validation.

The features were generated with the help of CMU Toolkit (for vocabulary and n-gram generation) and a CRF-based phrase chunker (for POS tagging and phrase chunking).

Finally, we do not present results for all possible combinations of feature characteristics. This is both because of limited space and because in some cases the combination is not very sensible. For example, we do not explore generalizations using POS and lexicons with single-word feature units because the resulting feature vectors will be somewhat trivial. And too given space restrictions we present results of such generalizations only with phrases and not n-grams. Similarly, we present results with feature selection strategies only for word based representations and not for

Table 4.1: Datasets

Name	**Reviews of**	**# of Neg.**	**# of Pos.**	**Total Size**
PangLee	movies	1,000	1,000	2,000
Jindal	products	2,520	17,480	20,000
Blitzer	products	16,576	21,972	38,648

n-gram and phrase representations. Presentation of results for such combinations are left to future research.

4.4 Results

Table 4.3 presents our first set of results. For each dataset we give classifier performance scores in terms of overall accuracy, and the F-measure (which combines information about both precision and recall) for negative and positive classes. Table 4.3 presents results using single words, stems, n-grams and basic phrases as the feature units, and Table 4.2 describes the approach of each run. The baseline used here is that of the majority vote where each document is assigned the dominant class.

4.4.1 Single-word features

4.4.1.1 Stemming

Although popular in information retrieval, stemming does not always add value in this task. By not stemming the terms in run 1, the accuracy improves on average, but insignificantly compared to run 2. Although the improvement is more pronounced for Pang & Lee dataset, with an increase of significance at $p = 0.055$ between runs 3

Table 4.2: Run design description

Run #	**Stemming**	**TF vs binary**	**Neg. words**	**n-gram**
1	no	TF	no	–
2	yes	TF	no	–
3	yes	bin	no	–
4	no	bin	no	–
5	no	TF	yes	–
6	no	TF	no	2
7	no	TF	no	3
8	no	TF	no	1,2
9	no	TF	no	1,2,3
10	no	TF	no	phrase
bl	majority rule			

and 4 (which are otherwise identical).

4.4.1.2 Term frequency versus binary weights

Comparing run 2 (TF) to run 3 (binary weights) as well as run 1 to run 4, we see insignificant changes in performance for all datasets. Similar to [60], we do not see a noticeable advantage of using binary instead of term frequency weighting. Note that there is, however a significant change in the F-measure for the negative class in Jindal dataset. Recall that this dataset is the most challenging as it contains only 12.6% negative documents, resulting in a lower classification performance for this under-represented class. Because the minority class is often of interest, features that help classifying it bears study in further research.

Table 4.3: Performance for single-word and n-gram features

Run	**Pang & Lee**			**Jindal**			**Blitzer**		
#	**Acc**	F_n	F_p	**Acc**	F_n	F_p	**Acc**	F_n	F_p
1	0.858	0.860	0.856	0.926	0.655	0.959	0.864	0.841	0.881
2	0.848	0.849	0.847	0.925	0.655	0.958	0.862	0.839	0.880
3	0.841	0.841	0.841	0.926	0.684	0.958	0.858	0.835	0.875
4	0.859	0.859	0.858	0.925	0.677	0.958	0.859	0.836	0.876
5	0.866	0.868	0.864	0.929	0.667	0.960	0.867	0.845	0.884
6	0.851	0.858	0.843	0.910	0.496	0.951	0.855	0.825	0.877
7	0.788	0.816	0.751	0.877	0.075	0.934	0.816	0.776	0.832
8	**0.875**	**0.879**	**0.869**	0.913	0.547	0.952	0.879	0.856	0.896
9	0.830	0.843	0.815	**0.947**	**0.748**	**0.970**	**0.896**	**0.876**	**0.910**
10	0.767	0.783	0.749	0.881	0.228	0.936	0.813	0.768	0.844
bl	0.500	0.500	0.500	0.779	0.126	0.874	0.510	0.430	0.570

4.4.1.3 Negations

Adding negated-word features in run 5 has proven to be marginally useful. Compared to otherwise identical run 1, the improvement has been made at insignificance levels for all of the three datasets.

Run 5 is the best performance we have achieved with words as the basic unit for all of the datasets. Note the accuracy achieved for the standard Pang & Lee dataset by this relatively simple document representation outperforms many approaches, including [54] and [42].

4.4.2 N-grams

Runs 6 through 9 include n-gram features of n up to 3. To test the effect of each level of n, all other aspects of the feature space were kept constant – stemming was not used, term frequency was used, and no negated-word features were added.

It is clear that the higher n-grams alone decrease the accuracy for all datasets. Run 8, which includes 1- and 2-gram features, performs the best for the smallest dataset, and run 9, which includes 1-, 2-, and 3-grams, is best for the other two. These results suggest that the n should be chosen appropriately for the size of the dataset. Perhaps the longer strings are not useful enough in smaller datasets. This supports the findings of [62] and [12], who reached different conclusions while working with datasets of different sizes.

4.4.3 Phrases

Run 10 shows the performance of the baseline phrase run wherein the classifier is trained on the phrases generated by a lexical phrase chunker. Although a vast improvement over majority baseline, the performance is significantly lower than single-words run 1 for all datasets. This result is surprising, in that semantically cohesive features intuitively should better represent the document. This feature type may warrant closer inspection in future work.

In summary, comparing words, n-grams and phrases we find that combining single-word vocabulary with n-grams proves to be the best strategy. Whether we use stemming, different word weighting or negation-enriched features did not prove to matter much in general, though some of these techniques may prove useful when datasets are small or have a vastly underrepresented class.

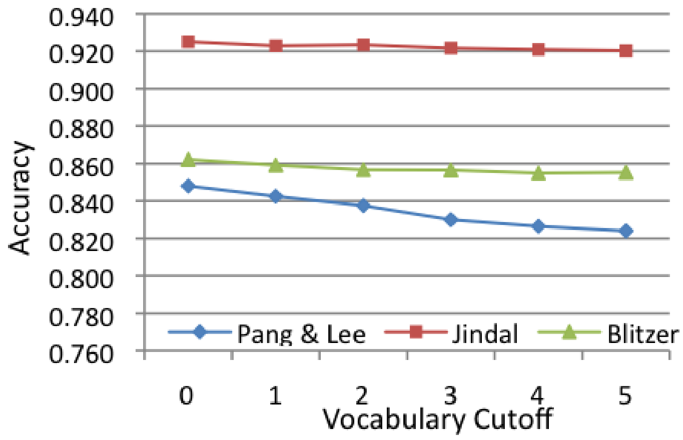
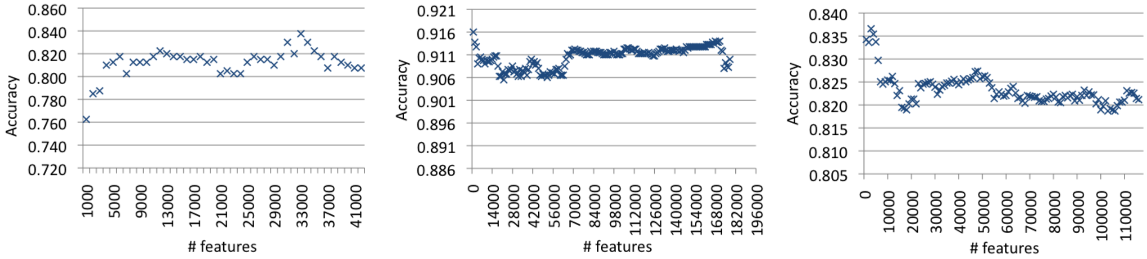


Figure 4.1: Feature selection of single-word runs using frequency-based vocabulary cut-offs

4.4.4 Feature Selection

4.4.4.1 Frequency-based Selection

In Figure 4.1 we explore the merits of cutting off the "tail" of the vocabulary, that is, excluding the terms that appear fewer than c times in the dataset from the feature space. The decrease in the performance compared to full-vocabulary run was not significant at $p < 0.05$ level up to $c = 3$ for Pang & Lee, $c = 4$ for Jindal, and $c = 1$ for Blitzer datasets (that is, when words appearing c times or less were excluded). This means that we can get an equivalent performance from a classifier for Jindal dataset while excluding words that appear 4 times or less in the dataset (leaving only 15.3% of original feature set!). Notice the differing acceptable cutoffs for the three datasets, which suggests that classification of some datasets is more sensitive to rare words than of others.



(a) Pang & Lee (b) Jindal (c) Blitzer

Figure 4.2: Performance with MI feature selection at various cut-offs

4.4.4.2 Mutual Information

To test the effect of Mutual Information feature selection on polarity classification performance, we divide each dataset into training (60%), tuning (20%), and testing (20%) subsets. For the two datasets in which the polarities do not have an equal share, the proportion of negative to positive documents was kept constant. The experiment was performed as follows. Features were extracted from the training set and their MI scores were calculated. The features were calculated using the settings of the best word-based run – run 5. After sorting the features in the ascending order, top N were chosen to represent the documents in the tuning set, with N varying from top few documents to the size of the feature space. Finally, for each dataset an N was chosen to maximize performance, and the testing set was used to determine classifier performance at this cutoff.

Probably because features are developed on a subset separate from the ones they then represent, the accuracy is negatively affected by the mismatch in vocabulary. This is especially evident in the smaller dataset (Pang & Lee) where the best accuracy

Table 4.4: POS and Lexicon-based feature selection for single-word features

	Pang & Lee		**Jindal**		**Blitzer**	
Run	**Acc**	**# features**	**Acc**	**# features**	**Acc**	**# features**
ADJ	0.781	13,546	0.901	21,150	0.772	16,217
VB	0.690	11,845	0.885	20,739	0.748	16,853
NN	0.756	26,965	0.882	84,510	0.758	60,034
ADJ $\cup$ VB $\cup$ NN	0.846	43,223	0.921	111,675	0.851	81,095
ACT	0.678	3997	0.902	3997	0.674	3997
SWN	0.819	52902	0.875	52902	0.797	52902
WNA	0.693	2367	0.876	2367	0.656	2367
run 1	0.858	50,917	0.926	218,103	0.864	153,789
majority	0.500	—	0.779	—	0.510	—

goes from 0.866 (run 5) to 0.838.

Figure 4.2 shows the performance of the classifiers at various cutoff points. For all datasets, the performance drops off as the number of features approaches 100% (the number of features is different for each dataset). This means that when sorted by MI, the bottom features hurt the performance of the classifier. Towards the top of the list, the performance differs between the relatively small Pang & Lee dataset and the others, which are larger by an order of magnitude. For the smaller dataset, using only top few thousand features hurts performance, but the best performance for Jindal and Blitzer datasets is achieved with only the top few thousand features. This suggests that dataset size influences the feature selection strategy and the thresholds to be used. We noted the best cutoff point for each dataset and use the testing set to get the accuracy scores of 0.798 (Pang & Lee) at 76% cutoff, 0.903 (Jindal) at 1%, and 0.837 (Blitzer) at 3%.

4.4.4.3 POS-based Selection

To see whether focusing only on certain parts of speech helps in polarity classification, we systematically excluded all but adjectives, verbs, or nouns. Results can be seen in Table 4.4. For each dataset besides accuracy we present the number of features for each run. Although the best accuracy is achieved when all three parts of speech are used, the best improvement attained per feature is with adjectives, and secondly with verbs, showing that these two parts of speech are indeed more helpful in polarity classification.

4.4.4.4 Lexicon-based Selection

Constraining the feature space to sentiment-annotated lexicons is another way to use external knowledge for feature selection. Second half of Table 4.4 shows the performance of the classifier trained on features limited to one of the three lexicons. The Affect Control Theory (ACT), SentiWordNet (SWN) and WordNet-Affect (WNA) lexicons contain 3997, 52902, and 2367 terms, respectively. The largest lexicon, SWN, provides the best performance for Pang & Lee and Blitzer datasets. Yet in Jindal its performance is equivalent to the WNA run, making its improvement/feature ratio 25 times less than that of the WNA run.

In summary, we find that using collection-specific measurements such as term frequency and MI we can successfully decrease the feature space, and in the case of MI significantly improve the performance. Parts of speech and lexicons did not prove to be as useful, suggesting that the vocabulary of the dataset is the best way to start building a classifier.

4.4.5 Feature Generalization

In attempt to assess the usefulness of POS and lexicons in another scenario, we performed feature generalization by applying several heuristics to phrases. Recall that the baseline phrase run 10 in Table 4.3 is outperformed by single-word and n-grams. Because of the large number of features generated in this test, only results for Pang & Lee are discussed in this section. First, we took the most drastic approach and replaced all words with their POS, attaining an accuracy of 0.627 (which is surprising, considering the triviality of resulting features). To avoid over-generalization we then replaced all words with POS except some words that we considered important to the task. When we consider the words in sentiment-annotated lexicons as important, we significantly improve performance, best of all with SWN lexicon to up to 0.754 accuracy. Alternatively, when considering various POS as important, we achieve the best performance of 0.742 with adjectives, again suggesting that they are indeed useful. In the next section, we explore the importance of these features compared to the rest described in this chapter, as determined using expected Mutual Information.

4.4.6 Feature Usefulness Across Feature Types

After generating all feature spaces for different feature types, we examined them in combination using expected MI discussed in MI Based Feature Selection Section. In particular, we wish to identify the most useful (highest MI) features irrespective of type.

Table 4.5: Top 50 features from Pang & Lee dataset selected using MI

bad	worst	stupid	boring
the worst	waste	ridiculous	wasted
awful	outstanding	mess	neg-even
life	waste of	of the best	lame
supposed	perfect	wonderfully	of the worst
should have	he is	memorable	supposed to
excellent	as the	dull	poorly
perfectly	both	subtle	script
plot	?	allows	performances
wonderful	also	bad movie	terrible
terrific	world	effective	the best
finest	hilarious	true	to work with
	ludicrous	boring	

Table 4.5 shows the top 50 features selected using MI for Pang & Lee dataset. All are either single-word feature or 2- or 3-gram. Many are obvious choices of opinion-laden words, such as *bad*, *worst* and *stupid*. But some are surprising, such as *both*, *?* and *as the*, suggesting that some stopwords and punctuation may be good indicators of sentiment polarity. Furthermore, we see features like *script*, *plot*, and *bad movie*, which are specific to the main topic of the dataset. Using MI to select features, we may be able to create not only a highly accurate, but also domain specific sentiment lexicons that cuts across feature types. Table 4.6 shows the number of various features that appear in the list of top 1000 features ranked using MI. Here, *Phrases* refers to the baseline phrase run and *Generalized phrases* refers to the best-performing generalized phrase run.

Table 4.6: Number of features by type in top 1000 from Pang & Lee dataset selected using MI

Feature type	**# out of 1000**
Single-word	330
Negation-enriched	13
2-grams	394
3-grams	138
Phrases	62
Generalized phrases	63

4.5 Cost Analysis

Finally, we analyze the computation time needed to generate the various features and the space needed to store them. The first two columns of Table 4.7 show the number of features and size of the standard Weka ARFF file containing them (in sparse format) for Pang & Lee dataset. The largest files produced by far were the n-grams, followed by phrases. The last two columns show the time (in milliseconds) it takes to generate the feature space and the average time it takes to generate a feature vector for each document. The tests were run on a computer with AMD Athlon 64 Processor with 1024KB cache and 1GB RAM. Although in terms of number of features negation-enriched features are few compared to the other types of features, because templates are used to extract these, the time it takes to generate the feature space is even greater than that of generating the 2-gram feature space. Also, the size of the lexicon used for generalization greatly affects the time it takes to generate the feature space as well as to process each document.

Table 4.7: Space and computation time statistics for various features for Pang & Lee dataset

	Space to store...		**Time to generate... (ms)**	
Feature type	**# of feats**	**space (bytes)**	**feat space**	**doc vector**
Single-word	50,918	6,513,249	5,917	584
Negation-enriched	2,305	143,923	7,519	244
2-grams	468,023	24,142,950	7,483	4,254
3-grams	1,044,171	41,152,199	11,245	8,625
Phrases	171,515	8,026,851	141,012	1,151
Gen-zed phrases (ADJ)	145,650	7,687,189	49,121	716
Gen-zed phrases (VB)	146,365	7,415,978	47,575	715
Gen-zed phrases (NN)	101,337	5,402,528	50,224	719
Gen-zed phrases (ACT)	156,907	7,441,326	69,953	736
Gen-zed phrases (SWN)	97,634	4,897,393	1,061,552	1,428
Gen-zed phrases (WNA)	164,631	7,763,846	97,972	717

4.6 Conclusion

4.6.1 Summary of Findings

In our exploration of some of the latest popular feature definition, selection, and generalization techniques, we use three datasets to test techniques popular in SA literature. We confirm some hypotheses, including that adjectives are important for polarity classification, and that stemming and using binary instead of term frequency feature vectors do not impact performance. We also show that the helpfulness of certain techniques depends on the nature of the dataset. For example, selecting top few features using Mutual Information hurts performance of the classifier on a smaller dataset, whereas it proves to be a good strategy for larger datasets.

Finally, we present the cost analysis in terms of space used to store the dataset and the time it takes to compute it. We see that, for example, it takes more time to compute negation-enriched features (using templates) than it takes to compute the whole vocabulary, putting in question any benefit these may give when working with large datasets.

Following the findings in this chapter, we use a 1,2,3-gram feature space with term frequency weighting in the following chapters.

CHAPTER 5
CROSS-TOPIC AND CROSS-STREAM SENTIMENT CLASSIFICATION

When we speak of *Social Media* today, we refer to a wide variety of social forums: blogs, wikis, review forums, social networking sites, and many others. However, SA research has not considered comparing sentiment *across* different social media. In this chapter, we compare sentiment over three particular media: Blogs, microblogs (Twitter) and Reviews (in the next, we will compare YouTube and Twitter). Additionally, instead of adopting the somewhat standard approach of analyzing documents on just a few select topics [6], we adopt a deeper topical perspective. Specifically, we compare sentiment across these three media as expressed on a common set of consumer product topics distributed under five topical categories (whereas in the next chapter we focus on political topics).

Our second major goal is to explore how best to build SA classifiers when we have data along a two-dimensional grid, one varying over source and the other on topic. In particular we use this data to explore stream (or source) adaptation and answer the question: *to what extent can a SA classifier developed on a topic for one social medium be transferred for use on a different medium?* There is related prior work, as for example, in [6], but instead of being cross source, they explore cross topic SA classifiers within the same medium. For example they train classifiers on documents about electronics and test on other documents about movies. In contrast, [64] do explore cross-source adaptation, but only from microblogs to reviews, and again without topical constraints on the datasets. The difference in our work is that we study cross-source (aka domain) classifiers for the same topic. We explore both single-source and multiple-source adaptation. For example, we study SA classifiers trained on reviews and blogs and tested on tweets for the same topic. We also study voting approaches as another angle for combining SA classification knowledge. We show, for example, that Twitter, despite its size restrictions, is a good source for building classifiers to be used in other kinds of data.

We first build a multi-dimensional dataset spanning the three social media sources for 37 topics distributed across five topical categories. We label sentiment in our dataset using Mechanical Turk taking suitable measures to ensure quality. After comparing sentiment across these streams, we use our dataset to evaluate classifiers designed to detect documents that express positive or negative sentiment. Furthermore, we conduct experiments to explore cross-stream as well as cross-topic classifier training and testing.

5.1 Dataset

We explore sentiment across three social media sources – blogs, microblogs (Twitter), and reviews. Each "stream" provides a somewhat different outlet for self-expression and discussion of various topics. However, this must be differentiated from a notion of synchronized streams often found in literature on emerging topic detection or topic tracking. Because of the data collection restrictions documents collected for each of these streams unavoidably cover a different time span.

Reviews are by definition almost tied to particular topics, for example a review may be on a specific game or a particular movie. Tweets are also, for the most part, bound to topic - though the topic itself may be indecipherable given that tweet language may be extremely casual. Blogs on the other hand may be focussed on a single topic or may be mixed and have themes that do not even relate to each other (e.g., personal issues interspersed with political opinions). Thus, we adopt unique approaches to collecting documents for each source.

Data collection in social media is itself challenging enough to be the subject of a separate study. This compounds when we try to maintain consistency across streams. Our procedure, described below, aims for roughly equal topic representations, i.e., retrieved sets in all three streams for each of our five topical categories. The categories are *movies*, *music albums*, *smart phones*, *computer games*, and *restaurants*. Each category consists of several topics (sometimes we call these queries), which were gathered from outside authoritative sources and pruned during data collection. The data collected was then cleaned and sampled. The blog and Twitter subsets were iteratively labeled for topicality and sentiment using Amazon Mechanical Turk. Reviews had their own in-built sentiment labels in the form of star ratings. The resulting datasets provide us with the opinionated texts on a controlled set of topics across three social media sources.

5.1.1 Data Collection

To get a list of most talked-about movies, we combined the Internet Movies Data Base (imdb.com) lists of all-time top movies with the listings of previous year and decade. The top 100 music albums on Amazon (amazon.com) provided recent as well as popular albums titles. Computer game topics came from a crawl of Metacritic (metacritic.com), a movie and game review website and included 9 platforms: PS3, Xbox 360, Wii, PSP, DS, 3DS, PC, PS2, iOS. The smart phone list came from a list of most popular phones for the major manufacturers on CNET (cnet.com). Finally, a list of restaurants came from a combination of the 20 most popular restaurants from 12 largest cities in the United States, as discussed on restaurant review website Yelp (yelp.com). For each topic, a query was designed to describe the topic in most unambiguous way. For example, the query "*Halo 3 Xbox*" contains both the name of the game and the console for which it is made. The same query was submitted to each stream. We take an iterative approach to choose the topics from our initial set. Starting at the top of each initial list, we retrieve documents using the following 2-part rule:

- if # of returned results from any stream is $< 50 \rightarrow$ discard the topic, else keep the topic
- if # of returned results from a stream is $> 100 \rightarrow$ select 100 randomly

The topics passing the above rules are retained in their topical category. We iterate through topics until we have retrieved a minimum of 500 documents in each stream per topical category. The final collection is then cleaned using stream-specific

Table 5.1: Dataset statistics

	#queries		**For chosen, # of docs**			**After cleaning & sampling**		
Category	**Tried**	**Chosen**	**Twitter**	**Reviews**	**Blogs**	**Twitter**	**Reviews**	**Blogs**
Movies	19	8	7538	17707	510	770	800	423
Music	56	8	5738	1760	556	740	772	467
Games	7	7	2193	2123	584	495	617	525
Phones	12	5	7479	2146	573	482	500	432
Rest-ts	26	9	881	22731	614	566	900	355
Total	120	37	23829	46467	2837	3053	3589	2202

approaches as described below. The number of topics explored and selected, the number of documents initially retrieved and finally the number of documents left after data cleaning and sampling is in Table 5.1. Computer games was the easiest topic to sample, with 7 out of 7 tried topics accepted. The most difficult one proved to be musical albums, with very little discussion about the older ones in the blog stream (recall that the initial list consisted of all-time top 100 albums). The blog stream was the limiting factor for most of the topics, often returning fewer than 50 documents. The final collection consisted of 8844 documents, 6642 of which needed to be labeled (reviews already had star ratings from which labels could be extracted). Table 5.2 shows the selected topics for each topical category, as well as the number of documents initially retrieved and remaining after cleaning and sampling. Notice that although Twitter often provides a good number of results, the sample after cleaning decreases dramatically due to a large amount of duplication.

Finally, Table 5.3 shows document length statistics for each topical category in the three streams. Notice that the maximum length of tweets exceeds the 140

Table 5.2: By-topic statistics

		# of docs retrieved			Cleaned & sampled		
#	**Topic**	**Twitter**	**Reviews**	**Blogs**	**Twitter**	**Reviews**	**Blogs**
	MOVIES						
1	The Dark Knight (2008)	1500	3301	60	100	100	37
2	The Godfather (1972)	1500	1286	64	100	100	51
3	Fight Club (1999)	1500	1953	51	100	100	41
4	The Lord of the Rings: The Fellowship of the Ring (2001)	582	4083	76	100	100	55
5	The Matrix (1999)	458	2544	76	100	100	66
6	Inception (2010)	1500	2003	63	100	100	57
7	American Beauty (1999)	72	1980	57	70	100	56
8	Star Wars: Episode V - The Empire Strikes Back (1980)	426	557	63	100	100	60
	MUSIC ALBUMS						
1	Barton Hollow - the Civil Wars	161	72	58	92	72	49
2	Wasting Light - Foo Fighters	556	223	61	100	100	52
3	21 - Adele	1401	360	75	100	100	62
4	The King is Dead - The Decemberists	173	102	65	91	100	56
5	The King of Limbs - Radiohead	370	189	82	100	100	59
6	So Beautiful or So What - Paul Simon	77	106	74	57	100	69
7	Back to Black - Amy Winehouse	1500	538	69	100	100	61
8	Teenage Dream - Katy Perry	1500	170	72	100	100	59
	COMPUTER GAMES						
1	Counter-Strike: Source (PC)	112	59	84	39	59	79
2	Half-Life: Counter-Strike (PC)	73	270	86	46	100	82
3	Call of Duty: Modern Warfare 2 (PC)	816	684	102	100	100	96
4	LittleBigPlanet (PlayStation 3)	83	300	77	83	100	53
5	Team Fortress (PC)	532	58	63	55	58	61
6	Left 4 Dead (PC)	206	230	84	72	100	73
7	Halo 3 (Xbox)	371	522	88	100	100	81
	SMART PHONES						
1	iPhone	1497	490	98	100	100	77
2	Motorola Droid	1500	577	95	82	100	80
3	HTC Evo 4G	1500	505	106	100	100	92
4	HTC Incredible	1500	356	175	100	100	100
5	HTC Thunderbolt	1482	218	99	100	100	83
	RESTAURANTS						
1	Tartine Bakery - San Francisco	62	3103	65	51	100	65
2	Bottega Louie - Los Angeles	168	3072	86	100	100	25
3	Wurstkuche - Los Angeles	163	2766	86	100	100	30
4	Gary Danko - San Francisco	55	2749	52	30	100	49
5	House of Prime Rib - San Francisco	87	2502	38	77	100	36
6	Kuma's Corner - Chicago	60	2335	53	54	100	34
7	Shake Shack - New York	141	2287	78	30	100	75
8	Din Tai Fung Dumpling House - LA	72	2029	87	55	100	14
9	Griddle Cafe - Los Angeles	73	1888	69	69	100	27

Table 5.3: Document length statistics

		Words				Characters			
		Mean	**Stdev**	**Min**	**Max**	**Mean**	**Stdev**	**Min**	**Max**
Twitter	Movies	16.9	6.0	2	31	95.6	32.6	10	160
	Music	14.5	5.6	3	32	81.9	30.3	12	159
	Games	18.6	5.7	7	32	97.6	29.3	34	156
	Phones	17.3	7.0	3	33	93.6	36.5	17	161
	Restaurants	15.9	5.6	2	30	89.9	30.9	14	159
Reviews	Movies	233.2	194.1	13	1051	1349.9	1142.2	71	6199
	Music	149.8	168.3	12	1473	821.2	942.4	57	8122
	Games	176.6	246.0	8	2722	957.2	1383.5	33	14741
	Phones	181.5	186.7	8	1365	977.9	998.8	38	7437
	Restaurants	146.3	118.8	1	877	790.7	642.1	19	4921
Blogs	Movies	225.8	336.9	20	3496	1309.9	1965.3	152	20530
	Music	151.6	266.7	8	2767	874.0	1531.0	90	15884
	Games	184.3	268.6	6	2368	1056.7	1523.0	146	13277
	Phones	164.3	214.4	15	1522	945.0	1257.4	99	8949
	Restaurants	246.0	397.5	8	3544	1468.1	2509.4	96	23211

character limit, since Twitter allows “re-tweeting” by prepending “RT @user:” (where @user is the username of the author of the original tweet) without shortening the length of the original message. Review and blog streams have comparable mean document lengths in all topical categories (within 20 words) except for restaurants, where the difference is 100 words. Yet the two streams differ in standard deviations of the distributions, with blogs having a much more varied length. The distributions of lengths can be seen in Figure 5.1. We can see that restaurants have the most drastic difference between the mean document length of reviews and blogs.

5.1.1.1 Twitter and Blog Search

The number of the collected results was markedly different for each stream, illustrating differences in topic coverage in each. Using Twitter search API we retrieved up to 1500 tweets for each query. The returned tweets go up to two weeks

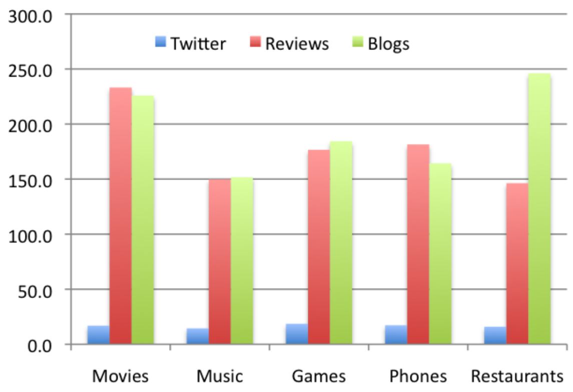


Figure 5.1: Distribution of mean document length in number of words

into the past. When sampling, tweets of fewer than 10 characters in length were not accepted. Duplicate documents were detected using textual content analysis. This is necessary, since for instance Twitter users may retweet a message without modifying it. For blogs we first used Google Blog Search API to retrieve 1000 results for a query. The pages these results linked to were then downloaded and the content was extracted. To extract the blog post content we first look for the title of the blog post (given by Google API), and analyze the text after it to get the content in which there are relatively few HTML tags. Specifically, as we process the text we keep track of a *tag density* measure, conveying to us how much text is shown compared to the number of HTML tags. We start collecting blog post content when there are five consecutive words without HTML tags and stop when tag density spikes. The gathered text also must consist of at least 90% alphanumeric characters, and must be at least 100 characters in length. These parameters were selected empirically.

5.1.1.2 Review Scraping

Instead of relying on a search interface, our approach to collecting reviews consisted of scraping various websites, thus providing a markedly cleaner dataset. To collect these reviews a scraper was coded for each website: IMDB.com for movies, CNet.com for phones, yelp.com for restaurants, and amazon.com for computer games and music albums. The main parts collected included star rating and review text, where the star rating was used as an indication of sentiment polarity. The reviews were randomly sampled to produce the final set for classification.

5.1.2 Data Labeling

We use Amazon Mechanical Turk[1] (AMT) to obtain labels for the blog and Twitter subsets. Providing a marketplace for work that requires human intelligence, such as data labeling, AMT has become popular in information retrieval and machine learning research [74]. However, rife with bots with some users not providing quality work, data gathered on AMT must be cleaned and quality control set in place.

Aiming to collect three ratings for each document, we designed two tasks (or Human Intelligence Tasks – HITs), one for blogs and another for tweets. Only raters with approval ratings over 90% were allowed to participate in the task. Each blog HIT contained 5 blog posts and Twitter HIT 10 tweets. At the end of each task the annotator is asked to enter the first word of the last document in the HIT as a quality control measure. Within the 10 tweets we also insert a “control” tweet with an

[1] http://www.mturk.com/

obvious polarity. If either control is failed, the whole HIT is rejected. The tasks were published in stages. HITs rejected during the first stage of rating were re-published. Only two such stages were necessary to collect 99% of the desired HITs.

Raters were asked to annotate each document (blog post or tweet) first for topicality – whether the document is relevant to the query – with available choices being *Yes*, *No*, and *Can't Tell*. For topical documents the rater is asked to select what kind of sentiment it expresses toward the topic: *Positive*, *Negative*, *Mixed*, *None*, or *Can't Tell*. By allowing the rater the choice of *Mixed*, *None*, or *Can't Tell* instead of forcing a choice between the two polarities, we improve the quality of the ratings and of the resulting dataset. Annotation guidelines for both tasks can be found in Appendices B and C.

The task proved to be challenging to the raters. We calculate inter-annotator agreement using a technique designed specifically for AMT tasks [77]. The measure is calculated by averaging Pearson correlation for each set of ratings with the average rating. We analyzed the labeling process in three stages. First, annotators had to decide whether the document was on topic. The agreement on this task was 0.600 for blogs and 0.389 for Twitter. Next, the topical documents had to be rated according to their sentiment. The agreement on whether the document had sentiment (i.e. was subjective) was at 0.260 for blogs and 0.490 for Twitter. Finally, the task of distinguishing positive from negative documents had an agreement of 0.305 for blogs and 0.535 for Twitter. Blogs proved to be more challenging in sentiment classification task than Twitter, suggesting Twitter may be a more suitable data source for training sentiment classifiers.

To explore these difficulties further, we computed a conditional probability matrix for each choice, which can be seen in Tables 5.4 and 5.5. Each cell contains $P(column|row)$, for example, for Twitter $P(No|CT) = 0.248$. Annotators agree more if the diagonal values are large. For example, in blogs probability of marking a post *Can't Tell* is 0.819 if it's already marked *Can't Tell*. Small numbers in other cells indicate little confusion. For example, in Twitter it is very unlikely to see a tweet marked *Neg* if it's also labeled *Pos* ($P(Neg|Pos) = 0.014$).

For topical decision in Twitter we see a large $P(Yes|CT)$ term of 0.756 and $P(Yes|No) = 0.565$ indicating that many raters disagreed on whether a topically ambiguous tweet was topical. Though there was still a large amount of tweets which were unambiguously topical: $P(No|Yes) = 0.057$ and $P(CT|Yes) = 0.033$. Disagreement is somewhat less for blogs, with $P(Yes|CT) = 0.424$. The choice of *Can't Tell* turned out to be the least divisive decision, with raters agreeing on it far more than on *Yes* and *No*. For the sentiment classification task we see that raters more often disagreed on the positive class in blogs (see $P(Pos|Neg)$, $P(Pos|Mix)$, $P(Pos|Non)$, and $P(Pos|CT)$) than in Twitter. This distinction is not as apparent in $P(Neg|Pos)$, $P(Mix|Pos)$, etc because the number of positive documents is so large that these probabilities are very close to zero.

A subset – 10 Twitter and 10 Blog HITs – were rated by an expert not associated with the project, and ratings compared to the majority rating. A similar difficulty level was seen with 67.7% of Twitter and 58.0% of blog annotation overlap.

Table 5.4: Conditional Co-occurrence of Annotations

- Twitter

	Topicality			**Sentiment**				
	Yes	**No**	**CT**	**Pos**	**Neg**	**Mix**	**Non**	**CT**
Yes	0.352	0.057	0.033					
No	0.565	0.693	0.113					
CT	0.756	0.248	0.855					
Pos				0.492	0.014	0.025	0.228	0.107
Neg				0.123	0.634	0.135	0.269	0.205
Mix				0.227	0.178	0.776	0.363	0.266
Non				0.141	0.031	0.030	0.509	0.173
CT				0.202	0.064	0.068	0.528	0.764

Table 5.5: Conditional Co-occurrence of Annotations

- Blogs

	Topicality			**Sentiment**				
	Yes	**No**	**CT**	**Pos**	**Neg**	**Mix**	**Non**	**CT**
Yes	0.447	0.214	0.101					
No	0.156	0.434	0.102					
CT	0.424	0.564	0.819					
Pos				0.593	0.036	0.122	0.208	0.071
Neg				0.319	0.870	0.211	0.261	0.084
Mix				0.372	0.074	0.814	0.264	0.100
Non				0.249	0.034	0.104	0.744	0.083
CT				0.351	0.049	0.157	0.338	0.874

To further understand the kinds of difficulties raters were having we examined a sample of documents. As expected, many mismatches between the expert and the AMT annotators involved fine semantic distinctions of the task. Here are a few examples:

- Authorship of the opinion may be unclear. News reports which put music album in a favorable light may be tagged as *Pos*, even though the opinion is not necessarily that of the author. However, if one demands the opinion to come from the author, the appropriate tag should be *Non*:

  > *In the year that Adele has achieved both global fame and breathtaking sales figures, it's two other female British singer-songwriters whose very different but equally striking visions of their country burn brightest [...]*

- The target of the opinion may be unclear. For instance, the text below refers to the trailer of a computer game, though it can be interpreted as a judgment of the game itself.

  > *Bad trailer if you ask me I really cant imagine except for the widely embraced console support that this game will outsell Battlefield 3.*

- Several opinions and targets may be present. In the blog post below the author has a favorable opinion of the phone (*"prized"*), but an unfavorable one of the screen protectors. It is sometimes difficult to separate the relevant opinions from the ones on a given topic.

  > *First of all I have 3 evos in my family. This advertisement is for*

> *three screen protectors. I ordered it. [...] they make your 550 prized phone look cheap and broken. They are all different and always too big really really bad. [...]*

The fine distinctions between authorship and intended target of an opinion must be specified very clearly in the description of task. It is also possible that using a pool of untrained workers to label the dataset (as in AMT) requires more than three annotators to provide a quality majority decision. These observations should be taken into account in the future labeling efforts.

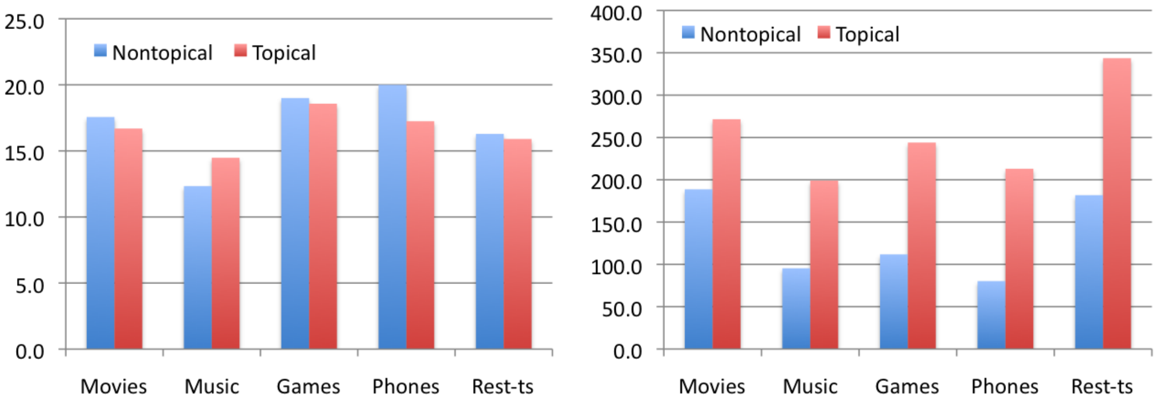
5.2 Stream Characteristics: Topicality and Sentiment

We now examine topicality and sentiment characteristics of the blog, Twitter, and review streams, as shown in Table 5.6 (summary). The final labels were decided using majority vote of the three ratings each document has received. Documents with no clear majority appear under Other. The column also includes entries marked *Can't tell*. The division between Pos, Neg, and Mix classes for reviews was done according to the star ratings. For five-star ratings we took 1-2 as Neg, 3 as Mix, and 4-5 as Pos. For ten-star ratings we took 1-3 as Neg, 4-7 as Mix and 8-10 as Pos. The percentages (in parentheses) for the topical classes are those of the total, and for sentiment classes are of the total number of topical documents.

Notice that Twitter generally returned larger numbers of documents of which a minimum of 80% were marked topical. For blogs on the other hand only 39 to 54% of the retrieved documents were topical. Intuitively, it makes sense that the longer

Table 5.6: Distribution of topical and sentiment documents (percentages are in parentheses).

	Blogs								
Category	Total	Topical	Not top.	Other	Pos	Neg	Mix	None	Oth
Movies	423	184 (44)	196 (46)	43 (10)	87 (47)	12 (7)	29 (16)	46 (25)	10 (5)
Music	462	243 (53)	160 (35)	59 (12)	154 (63)	8 (3)	19 (8)	51 (21)	11 (5)
Games	525	285 (54)	187 (36)	53 (10)	154 (54)	20 (7)	32 (11)	60 (21)	19 (7)
Phones	427	261 (61)	136 (32)	30 (7)	130 (50)	17 (7)	33 (13)	60 (23)	21 (8)
Rest-nts	355	138 (39)	172 (49)	45 (12)	96 (70)	2 (1)	17 (12)	15 (11)	8 (6)
Total	2192	1111 (51)	851 (39)	230 (10)	621 (56)	59 (5)	130 (12)	232 (21)	69 (6)
	Twitter								
Category	Total	Topical	Not top.	Other	Pos	Neg	Mix	None	Oth
Movies	770	612 (80)	126 (16)	32 (4)	182 (30)	41 (7)	16 (3)	319 (52)	54 (9)
Music	740	731 (99)	3 (0)	6 (1)	263 (36)	10 (1)	10 (1)	397 (54)	51 (7)
Games	495	473 (95)	14 (3)	8 (2)	128 (27)	26 (6)	42 (9)	231 (49)	46 (10)
Phones	482	479 (99)	1 (0)	2 (1)	187 (39)	99 (21)	29 (6)	142 (30)	22 (5)
Rest-nts	566	545 (96)	9 (2)	12 (2)	268 (49)	14 (3)	32 (6)	200 (37)	31 (6)
Total	3053	2840 (93)	153 (5)	60 (2)	1028 (36)	190 (7)	129 (5)	1289 (45)	204 (7)
	Reviews								
Category	Total	Topical	Not top.	Other	Pos	Neg	Mix	None	Oth
Movies	800	800 (100)	–	–	612 (77)	91 (11)	97 (12)	–	–
Music	772	772 (100)	–	–	627 (81)	84 (11)	61 (8)	–	–
Games	617	617 (100)	–	–	504 (82)	63 (10)	50 (8)	–	–
Phones	500	500 (100)	–	–	316 (63)	96 (19)	88 (18)	–	–
Rest-nts	900	900 (100)	–	–	715 (78)	70 (8)	115 (13)	–	–
Total	3589	3589 (100)	–	–	2774 (77)	404 (11)	411 (12)	–	–



(a) Twitter

(b) Blogs

Figure 5.2: Document length (in words) in topical and non-topical documents

documents such as blog posts would have more noise which would disrupt information retrieval. Our data distribution underlines the difficulty of retrieving topical blogs and the comparative ease of retrieving from Twitter. In terms of raw numbers too, Twitter appears to be a rich stream, supporting the recent wide use of Twitter for topic tracking [39, 16, 7].

Figure 5.2 shows length distributions of Twitter and blog documents for topical (documents with majority *Yes* on Topicality question) and non-topical documents (those with majority of *Not topical*, *Can't tell if it's topical*, as well as documents with no clear majority). Blogs show a significant difference between the lengths of non-topical and topical documents. Upon manual inspection, we find that the non-topical documents often are

1. related to topic but not about it (in this case, not about an album):

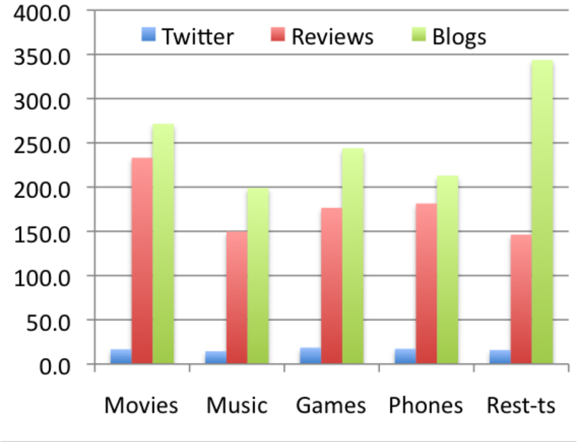


Figure 5.3: Document length (in words) in topical documents in the three streams

'Drunk' Amy Winehouse booed off stage in Belgrade By Entertainment in Video

2. selling or distributing a product as in

Album : 21 (Deluxe Edition) Year Of Release : 2011 Genre : Pop Quality : 320kbps Size : 155.78 [...] Code: http://www.wupload.com/file/64011602/ A21_320_FLAC.rar

3. related information, but not about the topic (in this case, music)

St. Vincent Announces Winter 2011 Tour Dates Next post: Neil Young Le Noise Documentary to Premiere at TIFF

4. about the author

I'm a student of Ancient History, Modern History and English. I

> *love music, art, movies and literature View all posts by adiek84*

5. non-content blog-related text

> *Enter your email address to subscribe to this blog and receive notifications of new posts by email.*

The relevant documents, on the other hand, are often long analysis of the topic, as for example a review of an album or a description of a trip. Thus, unlike in Twitter, topical blog posts tend to be longer than their non-topical counterparts. Figure 5.3 shows the distribution of document length in all three streams for topical documents. The difference between blogs and reviews is even more accented for restaurant topic, suggesting that people write about restaurants differently than about the other topics. A restaurant is more of an experience, and is often connected to a trip or an event, and thus may take longer to write about.

Upon examining the non-topical Twitter posts, we find posts similar to non-topical blogs. They are usually informative, and often promoting some product by posting a link. In fact, 51.2% of non-topical Twitter messages have a URL, compared to 43.3% of topical ones. The length distinction between topical and non-topical tweets is not as clear as in blog posts, since there is little range allowed in the first place. These observations may be useful for data collection and cleaning strategies for future dataset development endeavors.

Examining the topical documents we note that positive is the dominant class in all three streams, suggesting that overall sentiment trend is consistent throughout social media. However, there are some noteworthy differences. In Twitter the percentage of documents with no sentiment (category None) is typically double that of Blogs (except in Restaurants). This indicates that Twitter is not only a way for people to express their opinions, but is also a way to disseminate purely informational content. Furthermore, there is a non trivial portion of blogs which are of mixed sentiment, making this stream more challenging for sentiment analysis. Reviews, too, are not severely constrained in size, and thus allow for a more complex sentiment with 12 percent mixed documents.

There are also some topic-specific peculiarities which tell us about what kinds of topics people discuss on various social media streams. Phones show a large negative proportion of tweets and reviews, but not as many blogs, suggesting disgruntled electronics consumers revert to these social media to express their dissatisfaction. On the other hand, blogs provide a more complex discussion of movies, having Mix class more than twice as large as Neg, whereas Twitter provides very few mixed documents on the topic. Thus, sentiment extracted from each stream must be examined in the light of the stream's general tendencies about particular topics.

5.3 Sentiment Classification

As mentioned in the introduction, a part of our goal is to explore three different questions regarding sentiment classification. These concern classifying 1) different social media datasets when topic is controlled, 2) across different media and 3) across different topical categories. We design the classification task as follows. Given a relevant set of documents retrieved for a topic, identify the ones expressing positive and ones expressing negative sentiment towards it. This problem reflects the real-world situation where some documents may contain just positive, just negative, mixed, or even no sentiment at all. Compared to a binary positive/negative classification task popular in the sentiment analysis literature today [16, 62, 60], this task is more difficult, but makes fewer assumptions about the nature of sentiment expressions.

For each classification task – positive and negative – we build a distinct binary classifier and evaluate it using 3-fold cross-validation (choosing 3 folds in light of a small negative class). Note that having two binary classifiers allows us to place the mixed category in both relevant sets as a mixed-sentiment document is both positive and negative. For reviews we assigned one of the {*pos*, *neg*, *mix*} classes according to the star ratings as described before.

5.3.1 Within-Stream Sentiment Classification for Topics

In this section we study the question: how well do the current state of the art classifiers perform on different social media datasets when the topic set is controlled? We compare two classifiers: a Weka[2] implementation of SVM and a Lingpipe[3] language model-based logistic regression classifier, both of which have been used for sentiment classification [87, 17]. Whereas the Lingpipe classifier uses its language classification framework to extract features, we manually build features for the Weka SVM classifier. Choice of features come from our earlier preliminary study [51]. These

[2] http://www.cs.waikato.ac.nz/ml/weka/

[3] http://alias-i.com/lingpipe/

include:

- 1, 2, 3 grams extracted using CMU Statistical Language Modeling Toolkit[4]
- punctuation, separated using alphanumerical and whitespace characters
- counts of positive and negative smileys, as listed on Wikipedia List of Emoticons[5]
- count of negations such as "not" and "never"
- counts of various parts of speech, as extracted using Conditional Random Fields[6]
- number of URLs appearing in the text
- subjectivity/polarity priors from 3 lexicons SentiWordNet, WordNetAffect, and AffectControlTheory, computed as a number of positive/negative lexicon words found in the text

The results are shown in Table 5.7. For each task we show overall accuracy and the F-score for the sentiment class (positive for positive classifier and negative for negative classifier). F-score is the harmonic mean of precision and recall.

Considering individual topic categories, we note that *phones* were the most difficult category for Lingpipe to classify, despite a reasonable amount of negative sentiment present in the dataset. Furthermore, the extremely small negative representation in blogs on the topic of *restaurants* (2 documents) has made the negative

[4] http://www.speech.cs.cmu.edu/SLM_info.html

[5] http://en.wikipedia.org/wiki/List_of_emoticons

[6] http://crftagger.sourceforge.net/

classification extremely difficult, resulting in a 0.000 F-score for SVM. Finally, the scores for individual categories varied within each stream, suggesting that when one wishes to evaluate a classifier, a wide variety of topics should be included in the data set.

The task of classifying the under-represented negative class proves to be more challenging to the classifiers, which is especially evident in F-scores. Furthermore, Lingpipe classifier on average outperforms SVM, with the difference most drastically evident in the F-scores of the negative classifier. Lingpipe provides an average improvements in the minority class F-score of 6.8%, with a 13.3% improvement in the F-score of the negative classifier (one with underrepresented target class). The strongest performance overall is achieved using Lingpipe on reviews. This result makes sense, since reviews are written specifically to express sentiment. SVM classification of blogs showed the worst performance. Although positive F-scores for Twitter data are usually lower than those of other streams, negative F-scores are usually better. Thus, if one needs to find negative sentiment about a topic, Twitter is a good resource.

When performing an error analysis, we saw some regularities. The documents which are misclassified often fall into one of the following categories:

1. Comparisons. For example, the following is a review comparing *HTC Thunderbolt* phone to *iPhone*:

   > *The following is a list it's limitations (compared to iPhone): weaker battery, touch screen not as responsive or super sensitive, Apps are fewer and not as well screened (e.g., enough with the Porn), heavier,*

Table 5.7: Polarity classification (TW: Twitter, BL: Blogs, RV: Reviews)

	Positive polarity classification											
	Accuracy						**FScore**					
	LingPipe			**SVM**			**LingPipe**			**SVM**		
	TW	BL	RV	TW	BL	RV	TW	BL	RV	TW	BL	RV
Movies	0.706	0.656	0.883	0.694	0.658	0.888	0.581	0.783	0.941	0.585	0.780	0.940
Music	0.787	0.763	0.900	0.743	0.687	0.890	0.714	0.858	0.949	0.693	0.813	0.942
Games	0.709	0.632	0.898	0.710	0.628	0.896	0.608	0.763	0.948	0.607	0.770	0.945
Phones	0.610	0.636	0.815	0.534	0.617	0.804	0.561	0.749	0.906	0.604	0.719	0.891
Rest-nts	0.696	0.800	0.923	0.644	0.812	0.922	0.742	0.894	0.961	0.751	0.896	0.960
Avg	0.702	0.697	0.884	0.665	0.680	0.880	0.641	0.809	0.941	0.648	0.796	0.936
	Negative polarity classification											
	Accuracy						**FScore**					
	LingPipe			**SVM**			**LingPipe**			**SVM**		
	TW	BL	RV	TW	BL	RV	TW	BL	RV	TW	BL	RV
Movies	0.905	0.778	0.802	0.918	0.772	0.781	0.357	0.096	0.483	0.324	0.000	0.138
Music	0.978	0.889	0.829	0.971	0.893	0.823	0.333	0.185	0.276	0.160	0.071	0.180
Games	0.860	0.816	0.815	0.867	0.804	0.812	0.383	0.302	0.276	0.198	0.097	0.094
Phones	0.704	0.775	0.699	0.729	0.805	0.684	0.289	0.038	0.514	0.431	0.038	0.282
Rest-nts	0.921	0.867	0.803	0.932	0.855	0.801	0.418	0.222	0.354	0.431	0.000	0.091
Avg	0.874	0.825	0.789	0.883	0.826	0.780	0.356	0.169	0.381	0.309	0.041	0.157

> *poorer multimedia.*

Classifier has mislabeled it into a negative class, even though the final verdict was positive:

> *But to me a reliable network is the first priority. So, it's not bad and fairly close to the gold standard*

2. Sentiment-laden subject. This is especially apparent in the topic of movies. A movie may be about death and killing (of negative polarity), but the author may still be writing positively about it. For example, the positive review of *The Godfather* may be overwhelmed by the discussion of the negative-polarity content of the movie:

> *Considering the fact that the movie is about the life of a crime family, the enormous amount of killing that there is throughout the film isn't too suprising. Maybe a little disturbing, but not suprisng. [...] "The Godfather" was a very superb movie [...]*

3. Frequent negations. Some writers use negations so much and so skillfully that it is difficult for a classifier to capture the reversal in polarity. For example, when an author describes dashed expectations about a restaurant, many positive words may be used:

   > *I don't know why everyone loves their food so much...*

4. Switching the subject. Similar to comparisons, when an author chooses to talk about another subject, the text no longer relates to the target. For example, talking about a computer game *Counter-Strike: Source*, this author recommends *Half Life 2* instead:

   > *Simply put, you can buy half life 2, which is cheaper, and get 3 of these games. Still great games, but if you like them, they all started out as mods for half life. Just buy HL2.*

5. Mixed sentiment. Automatic weighting of pro's and con's proves to be challenging to the classifiers. An important positive comment may be "overpowered" by the negatives, such as in this game review:

   > *[...] campaign rating 0/10 stars for me. Very Disappointed. Not as fun as the original COD. [...] Its a great game to rent and buy [...]*

Literature contains some approaches to tackling some of the above problems. For instance, [33] use sequential rules to find *comparative sentences* in order to extract comparative relations between entities. Negations have been studied by [13], who explicitly include the negation in the document representation by appending them to the terms that are close to negations. The exploration of techniques to address these issues we leave for future research.

5.3.2 Cross-Stream Sentiment Classification for Topics

Next, we address our second question, which is: how well do classifiers trained on one stream perform while classifying data from another stream? This scenario is encountered when training data for a topic is only available from a different stream. To test cross-stream performance of our classifiers, we perform an evaluation using 3-fold cross-validation as in the previous experiment, except for using testing data of a different stream. The following experiments were done using Lingpipe since it gave better results in the previous experiments.

Our results are in Table 5.8. In each cell a classifier was trained on the data specified in the column and tested on data of the row ("target") stream. When the source for building the classifier differs from the target stream we refer to the classifier as 'foreign' and otherwise we refer to it as 'native' (the native runs are underlined). We present both Accuracy and target class F-score as measures. The best performance amongst the three streams for a given target-category-measure combination is in bold. When considering accuracy we see that the best performance within a category

Table 5.8: Single-Source Model Adaptation

			Accuracy			**Target F-score**		
Task	**Categ.**	**Target**	**Blogs**	**Reviews**	**Twitter**	**Blogs**	**Reviews**	**Twitter**
POS	Games	Blogs	0.631	**0.652†**	0.543	0.763	**0.824†**	0.627
		Reviews	0.788	**0.897**	0.865	0.880	**0.948**	0.927
		Twitter	0.528	0.365	**0.709**	0.520*	**0.646***	0.608*
	Movies	Blogs	**0.655**	0.638*	0.516*	0.783	**0.790***	0.568
		Reviews	0.731	**0.883**	0.759*	0.844	**0.941**	0.861*
		Twitter	0.534	0.367	**0.705**	0.439	**0.612***	0.581
	Music	Blogs	**0.762**	0.708	0.754*	**0.857**	0.848*	0.844*
		Reviews	0.856*	**0.900**	0.878	0.923*	**0.949**	0.937
		Twitter	0.670	0.380	**0.787**	0.558*	0.668*	**0.714**
	Phones	Blogs	**0.635**	0.608*	0.414	0.748	**0.792***	0.410
		Reviews	0.769	**0.815**	0.279	0.869	**0.905**	0.465
		Twitter	0.570*	0.513	**0.610**	0.531*	**0.637†**	0.561
	Rest-nts	Blogs	0.800	**0.814***	**0.814***	0.893	**0.905***	0.896*
		Reviews	0.882*	**0.923**	0.917*	0.938*	**0.961**	0.958*
		Twitter	0.548	0.539	**0.696**	0.651*	**0.753***	0.741
NEG	Games	Blogs	**0.815**	0.794*	0.805*	**0.302**	0.140*	0.126*
		Reviews	0.783	**0.814**	0.809*	0.103	**0.275**	0.037
		Twitter	0.721	0.838*	**0.859**	0.181	0.119	**0.383**
	Movies	Blogs	0.777	0.672	**0.783***	0.096	**0.240†**	0.179*
		Reviews	0.736	**0.802**	0.761*	0.192*	**0.483**	0.236*
		Twitter	0.844	0.800	**0.905**	0.282*	0.185	**0.356**
	Music	Blogs	**0.888**	0.810	0.884*	**0.185**	0.088*	0.000*
		Reviews	0.796	**0.828**	0.811	0.148	**0.435**	0.000
		Twitter	0.953	0.775	**0.978**	0.000*	0.152*	**0.333**
	Phones	Blogs	0.775	0.689	**0.814†**	0.038	**0.250***	0.197*
		Reviews	0.616	**0.698**	0.622	0.141	**0.513**	0.294
		Twitter	0.637*	0.578	**0.704**	0.226*	**0.440***	0.288
	Rest-nts	Blogs	**0.866**	0.837*	0.859*	**0.222**	0.000*	0.000*
		Reviews	0.764*	**0.802**	0.797*	0.186*	**0.354**	0.116*
		Twitter	0.863	0.874*	**0.920**	0.000	0.129*	**0.418**

- stream combination is mostly achieved by a native classifier. Specifically out of 30 accuracy measurements with native classifiers (5 topical categories * 3 streams * 2 classifiers) 26 native classifiers achieved the highest score. With Target F-score (negative class F-score for negative classifier, positive otherwise), which is the more challenging measure, fewer, i.e., only 19 native classifiers achieved the highest score. Overall these results with native classifiers are not surprising.

In contrast, the results look remarkably more interesting when we test differences in performance for statistical significance. We notice that there are many instances in which the performance of a foreign classifier is *statistically indistinguishable* from that of the native classifier – these instances are marked with a *[7]. For example, POS *restaurant* classifiers trained on reviews or on blog posts perform the same (in terms of Accuracy) as the corresponding native classifier. In some cases, as for example the POS *games* classifier trained on reviews even outperforms the native blog-based classifier at a statistical significance of $p < 0.01$! The four classifiers that significantly outperform their native counterparts are marked with †. The number of foreign classifiers which achieve performance statistically indistinguishable from or better than the native classifier (in 43 experiments out of 60) shows that cross-stream adaptation is possible, and in a few cases even beneficial. Thus the answer to our question is that we can, in general, build classifiers on one stream and use it on another. This facility is useful when it is hard to obtain sufficient topical documents

[7] In contrast to usual practice given our interest we mark the statistically indistinguishable results.

in a stream or it is challenging to label documents of a stream. We know from the previous section that blogs were more challenging to label than tweets both in terms of whether they carried sentiment and whether the sentiment was positive or negative. Our cross-stream results indicate that we could use data from other streams to classify blogs.

Table 5.9: Single-source model adaptation: number of best or statistically indistinguishable from best runs.

		Accuracy		**F-score**		**Either**
Source	**Target**	**NEG**	**POS**	**NEG**	**POS**	**All**
Blogs	Blogs	4	3	3	1	7
	Reviews	1	2	2	2	4
	Twitter	1	1	3	4	7
Reviews	Blogs	2	4	5	5	10
	Reviews	5	5	5	5	10
	Twitter	2	0	3	5	9
Twitter	Blogs	5	3	5	2	8
	Reviews	3	2	2	2	5
	Twitter	5	5	4	3	10
	Best possible	5	5	5	5	10

Examining Table 5.8 further we observe that if we total the number of times a stream offers the best score or a score that is statistically indistinguishable from the best (in accuracy or the target F-score) then we have the distribution as shown in Table 5.9. For example, the Blog stream positive classifier offers the best or similar to the best Accuracy in 6 out of 15 experiments (3 target streams * 5 topics). Of these 6, in 3 instances the classifier is a foreign classifier (i.e., classifying reviews or tweets).

It is not surprising to note that reviews are the best stream achieving a total of 29 instances across classifiers and measures with the best or close enough to best performance. Of these, 19 instances are in the role of foreign classifiers. What is most surprising is that Twitter is also a good source of training data, with best or close to best performance in 23 instances and these include 13 instances where the classifier is a foreign classifier. Blogs, on the other hand, offer the best or close enough scores in only 18 instances of which only 11 are as foreign classifiers. Moreover, when blogs or twitter posts are being classified, the native blog classifier is matched by a foreign classifier in 100% of the instances (compared to 60% for reviews).

Thus we infer, within the limits of these experiments, that blogs offer the least interesting classifiers for sentiment, whereas review-based classifiers are the best. Review classifiers offer the best or close enough scores in 10 out of 10 blog classification instances and 9 out of 10 Twitter ones, suggesting that tweets may be slightly more difficult to classify than blog posts. Surprisingly, this is followed by classifiers built on Twitter – a medium that is by design highly constrained in size. To further illustrate Twitter's strength, it offers the best or close enough classifier 5 times out of 10 even while classifying reviews and 8 out of 10 while classifying blogs.

5.3.3 Multiple-Stream Model Adaptation

We further explore cross-stream adaptation by taking advantage of several streams when building a classifier. The question we address is, *does training on several* *social media sources improve classification performance when adapting to another source?* We explore three scenarios:

- **Two-Source Mixed Model** – a classifier which has been trained on documents from two different streams (excluding target stream)
- **Three-Source Mixed Model** – a classifier which has been trained on three streams (including target stream)
- **Three-Source Voting Model** – three classifiers, each trained on one stream, using majority voting to determine the class of a document

An advantage of using several sources to build sentiment classifiers is the diversity of language and expression the training data includes, compared to training on only one source. We evaluate the performance of these classifiers as in the single-stream experiments above, using 3-fold cross-validation, and compare them to their native counterparts. For each of the above scenarios, ten experiments were conducted: one for each of the five topical categories, times two classifiers (positive and negative).

Figure 5.4 shows the accuracy of native, 2-source and 3-source mixed, and voting models with 99% confidence intervals. The results are presented first sorted by task (negative and positive), target stream, and finally by topical category. For example, the first interval shows the NEG classifiers used to classify documents on games in the Blog stream. We see that in some instances models match each other very well, such as when detecting negative blog posts (leftmost five tasks). Performances are not as much matched when classifying Twitter, though, especially with 2-source mixed model lagging behind the others. Notably, out of all of these experiments,

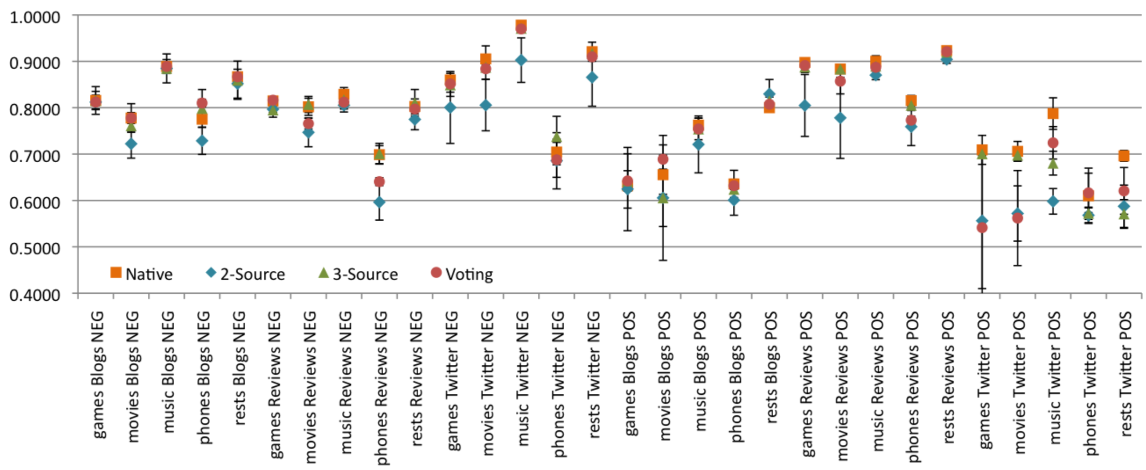


Figure 5.4: Accuracy of native, 2- and 3-source mixed,

and voting models with 99% confidence intervals.

only in one instance do we get a performance that is statistically better than that of the native classifier – the negative voting model tested on blog posts about phones. Otherwise, the performance is as good as or inferior to the native classifier.

Furthermore, we examine the number of best runs for each model in Table 5.10. We see that reviews benefit the least from a 2-mixed source model, followed by Twitter. Once again, blogs are shown to be easiest to classify using foreign training data with a matched performance in 10 out of 10 experiments for all models. Models which used all sources (mixed and voting) perform better than those which exclude the target data. This supports the common intuition that it is always beneficial to train on labeled data for the target dataset whenever possible. Looking closer at the distinction between the voting and mixed models when training on all three streams, we find that the mixed model predicts document class correctly 79.82% of the time

Table 5.10: Multiple-source model adaptation: number of runs statistically indistinguishable from native classifier.

		Accuracy		**F-score**		**Either**
Source Model	**Target**	**NEG**	**POS**	**NEG**	**POS**	**All**
Mixed R + T	Blogs	3	5	5	5	10
Mixed B + T	Reviews	1	0	2	0	2
Mixed B + R	Twitter	3	2	4	4	6
Mixed all	Blogs	5	5	5	4	10
Mixed all	Reviews	5	4	5	5	10
Mixed all	Twitter	5	3	5	1	8
Voting all	Blogs	4	5	5	4	10
Voting all	Reviews	5	3	4	4	9
Voting all	Twitter	5	5	2	5	10
Best possible		5	5	5	5	10

compared to 78.57% for the voting model, making the mixed model marginally better.

We conclude that compared to single-source adaptation, it is indeed better to train on many data sources as possible, that is, training on several different sources makes models more comparable to the native model. However, these models may only be comparable to the native model but they will not outperform it. Furthermore, mixing outside data with target data for training classifiers may produce weaker classifiers than if only the target data was used.

5.4 Topic-independent Experiments

Finally, to determine the influence of topic specificity on stream adaptation, we perform topic-independent experiments by combining the data across topics. The single-source, multi-source and voting model performances are shown in Table 5.11. Consistent with our earlier conclusions, we see that the best performance (in bold) is

Table 5.11: Topic-independent source adaptation results (native classifiers are underlined, best in **bold**, same as native marked with *, better than native marked with †)

		Accuracy					
		Single-source			**Mixed**		
Classifier	**Target**	**Blogs**	**Reviews**	**Twitter**	**2 source**	**3 source**	**Voting**
NEG	Blogs	**0.817**	0.732	0.773	0.788	0.801*	0.804*
	Reviews	0.662	**0.768**	0.642	0.636	0.735	0.715
	Twitter	0.791	0.762	0.862	0.816	**0.883†**	0.852*
POS	Blogs	0.628	0.683†	0.623*	0.660*	0.659*	**0.688†**
	Reviews	0.790	**0.881**	0.873*	0.821	0.881*	0.878*
	Twitter	0.620	0.448	**0.692**	0.631	0.654	0.655

		Target F-score					
		Single-source			**Mixed**		
Classifier	**Target**	**Blogs**	**Reviews**	**Twitter**	**2 source**	**3 source**	**Voting**
NEG	Blogs	0.232	0.282*	**0.325†**	0.202*	0.273*	0.256*
	Reviews	0.230	0.446	0.434*	0.304	**0.522†**	0.396
	Twitter	0.141	0.311	**0.450**	0.251	0.354	0.352
POS	Blogs	0.743	**0.835†**	0.721*	0.752*	0.747*	0.811†
	Reviews	0.880*	0.938	0.934*	0.901	**0.939***	0.937*
	Twitter	0.565*	**0.669***	0.668	0.551	0.484	0.652*

usually achieved either by the native model, or model using all three sources (3-source mixed or voting). However, the benefits of adapted models are not as pronounced as with topic-specific classifiers. For instance, the accuracy of negative classifiers targeting reviews is not matched by any adapted models. This is not true for topic-specific ones, with 3-source mixed models matching native classifier for all individual topics. The same is true for the positive classifiers targeting Twitter.

It is curious, however, that in some of the topic-independent experiments the foreign models significantly outperform the native ones, such as in the case of negative classifiers targeting Twitter (in Accuracy) and targeting reviews (in F-score). Thus, for topically-mixed collections, it is the case that information from a variety of topics from several sources may improve native classifiers. This was not the case for topic-specific experiments earlier, with only one of the multi-source experiments outperforming their native counterparts. We conclude, then, that it is not only beneficial to combine sources of data, but also the topical domains.

5.5 Cross-Topic Adaptation

Here we address our third question: how do sentiment classifiers when trained on one topic category perform on documents of a different category? Note that this time we constrain the training and testing sets to the same stream and vary only the topic category. For each topic category within a stream we broke the data into 3 folds. Taking 2 folds as training we build a classifier and test it on the 3rd fold. We repeat this three times and compute averages in performance. We refer to the classifier in this design as a “homogenous” classifier (training and testing documents are in the same category). Note this is essentially the same classifier as used in section 4.1. We then build a second classifier that is “heterogenous” in nature. Specifically it is trained on documents from the remaining four topic categories, i.e., excluding the topic category for the test documents. For this we randomly sample equal numbers of documents from the other topic categories such that the final training data size is the same as for the homogenous classifier. We then tested this “heterogeneous” classifier on the target test set. To minimize sampling bias, we build 10 heterogeneous

Table 5.12: Cross-topical classification (F-score)

		Twitter		**Blogs**		**Reviews**	
Task	**Target**	**Native**	**Others**	**Native**	**Others**	**Native**	**Others**
POS	movies	0.580	0.523 (9.8)	0.782	0.780 (0.3)	0.941	0.921 (2.1)
	music	0.714	0.536 (24.9)	0.857	0.801 (6.6)	0.949	0.940 (0.9)
	games	0.608	0.494 (18.7)	0.762	0.785 (-3.0)	0.948	0.945 (0.3)
	phones	0.561	0.633 (-12.8)	0.748	0.736 (1.6)	0.905	0.902 (0.4)
	rests	0.740	0.573 (22.5)	0.893	0.864 (3.2)	0.961	0.833 (13.3)
NEG	movies	0.364	0.161 (55.7)	0.179	0.065 (63.6)	0.485	0.263 (45.8)
	music	0.444	0.068 (84.6)	0.268	0.064 (76.0)	0.428	0.264 (38.3)
	games	0.371	0.177 (52.3)	0.302	0.163 (45.7)	0.276	0.198 (28.2)
	phones	0.291	0.183 (37.1)	0.052	0.085 (-64.0)	0.509	0.278 (45.3)
	rests	0.406	0.074 (81.8)	0.305	0.044 (85.6)	0.353	0.379 (-7.2)

classifiers and average over their performance.

We report only on F-scores which is our more challenging measure. These are in Table 5.12. The percentage changes when using the heterogenous classifiers are given in the parentheses so a positive percentage indicates a drop in performance. Due to space restrictions we do not show accuracy scores for this experiment.

Looking at the results in Table 5.12 we find that overall in almost all instances the heterogenous classifier did not perform as well as the homogenous one. However, there are some stream-specific trends. Especially with reviews and blogs within POS classifiers drops in performance were less than 5%. In four instances there were actually improvements when using the heterogenous classifier. Losses with heterogenous classifiers were considerably higher with the NEG classifiers than with the POS classifiers. Even for blogs and reviews we find losses between 28% to as high as 87%. Overall with a few exceptions we infer, within the constraints of our experiment, that homogenous classifiers are preferred to heterogenous ones. The penalty paid is greater when we try to classify NEG sentiment than when we aim to classify POS sentiment. In fact with blogs and reviews, it appears as if heterogenous POS classifiers offer somewhat comparable performance to homogenous ones, barring one or two exceptions. Twitter on the other hand seems difficult to classify if one moves away from the topic category for training documents. This is especially so for classifying NEG sentiment. This implies that Twitter uses a more topically unique language which is difficult to learn from other topics. With the other two streams the cross topic effectiveness is observable for POS classifiers but not for NEG classifiers.

5.6 Discussion

Finding labeled data to train classifiers can be difficult and expensive. During the creation of the dataset it was clear that some streams are easier to sample, collect, and label than others. For example, besides writing specialized site scrapers to collect reviews, very little post-processing needed to be done, and because of star ratings, no labeling. Using Amazon Mechanical Turk and paying as little as 0.5 cents per annotated document, we spent almost $200 labeling two out of three sources of data, getting just enough to claim statistical significance of our results. On industrial scale, labeled data may be not only expensive, but impossible to get, with the plethora of social media websites perpetually expanding the way people express their opinions.

The experiments described in this work demonstrate the effectiveness of using sentiment classifiers trained on one data source and applied to another. Not only are models trained on single sources often an adequate substitute for the native classifier, but in combination they are even more helpful, often performing as well as the native classifiers. It is interesting that the dataset which was the most challenging to collect and label – the blog stream – was most amenable to classifiers built from other sources. And the dataset which was the least challenging to gather and which did not even need human labeling – the review stream – proved to be the best source of training material for the classification models. It may be the case that the quality of data reviews provide, as well as unambiguous purpose of reviews (that is, to express opinions), overshadow any special language and style features of the other streams. These results are also in agreement with [4] who find that blogs are the most difficult to classify, followed by microblogs (such as Twitter), and the best classification performance is achieved by models trained on reviews (though note that their work was not on cross-stream classifier experiments).

On the other hand, other streams are not to be discarded in favor of reviews. Twitter is our second best source of training data. Unlike reviews, though, it is a much more topically diverse source of data. If one plans on classifying documents about products and services, reviews would be very helpful in building a classifier. But if one is interested in matters outside popular review websites – global issues in policy and economics, or personal ones like self-esteem or social anxiety – reviews may be of little help. It would be interesting to create a multi-dimensional dataset similar to the one in this study, but centered around topical categories not found on popular review websites. Thus, in the next chapter we examine political topics like politicians, issues, and events in two sources: YouTube and Twitter.

Another peculiarity of our dataset is the choice of popular topics for each of our categories. Not only are these topics more likely to be discussed and strongly represented in several social media streams, they are more likely to be discussed favorably, favoring the positive sentiment as seen in the statistics of our dataset. Although it is a perfectly reasonable way of selecting topics (they are popular, after all), the resulting dataset displays a class imbalance which presents difficulty in training classifiers. Notably, the political dataset described in the next chapter shows the opposite tendency to the negative sentiment. Finally, the data collection strategy we adopted, which includes a duplicate detection step, limits the data analysis since, as we show in the following studies, social media such as Twitter contain a fair amount of redundancy. In the studies described in the next two chapters we adjust our data collection strategies to keep the duplicate data.

5.7 Conclusion

In this study we create a multi-dimensional dataset in which three social media sources are queried using a common set of topics and we examine the differences and similarities of the sentiment expressed in these data streams. We then perform a series of experiments testing performance of models trained on annotated data within and across streams.

5.7.1 Summary of Findings

We conclude that although the general proportion of positive to negative remained similar across streams, there are marked differences between them. Our data indicates that stream-specific topical tendencies must be taken into account when mining sentiment.

Our stream adaptation experiments show the usefulness of each stream as a source of training data. Classifiers built using reviews prove to be the most generalizable to other streams, followed by Twitter, with Twitter-based model performing as well as the native classifier 8 out of 10 for blogs and 5 out of 10 for reviews. We also show that combining training data from several streams further boosts performance, and combining data from different topics may even produce classifiers outperforming their native counterparts.

Our study of the relative usefulness of social media streams as sources of training data allows for more informed design of sentiment analysis tools wherein resources are spent on collecting and labeling data best suited for the task.

5.7.2 Future Work

In these experiments we examine topics which were best suitable for mining these three streams, especially constrained to topics for which reviews can be found. If not restrained by such considerations, a wide variety of topics can be examined. These include current news, disasters, personal reflections, or even stock speculations. Furthermore, social streams are not limited to text – streams of video, images, and audio are available for multi-media analysis. For example, are sentiments expressed about physical exercise on youtube workout channels different from blog posts on personal webpages? Do videos provoke different emotions than images or text? It is now possible to address such questions using commentary provided popular commenting feature on most social media websites.

CHAPTER 6
POLITICAL SPEECH IN SOCIAL MEDIA STREAMS: YOUTUBE COMMENTS AND TWITTER POSTS

We continue our examination of sentiment across social media streams, this time focusing on political discourse. Originally dealing largely with product reviews [62], social media-driven sentiment analysis (SA) has recently expanded its target to encompass political discourse. We see also political sentiment research being framed as an analysis of the author's political *stance* (i.e. attitude adopted with respect to an issue) [78, 43], diverging from the standard practice of focussing on sentiment defined as (*positive* or *negative*).

Again, a limitation of such research both in the general sphere and in political arena is that the focus is on a single source at a time [7, 19, 22, 47, 84]. For example, Bollen et al. [7] use Twitter to estimate "public mood state". Would the same observations be made if more social forums were taken into the account? Livne et al. [47] use only Twitter to study the behavior of political parties in the 2010 election. Would this behavior look different if more sources were examined? Furthermore, when multiple sources *are* compared, such as news and blogs in [44] or reviews and blogs in [4], the topics considered in the sources differ, thus limiting the observations and conclusions that can be made. To the best of our knowledge there have been no efforts at comparing sources using the same set of topics.

The concern we raise about single-source analysis is important given the differences between social media sites. For example, Twitter is known for its 140 character post limit – potentially encouraging one-sided discourse instead of exploring several sides of an issue. Other sites focus on specific media, such as video (YouTube), pictures (Flickr), or bookmarks (Delicious). Driven by, for example, videos from a political rally, would discussion on YouTube be more animated or have more flaming than one on Twitter? Finally, privacy settings may affect the quality of content, with anonymous writings potentially differing from those with a real name, an effect known as a *online disinhibition* [82].

Given the above observations, our goal is to compare two social media sources – Twitter and YouTube. We focus on textual content and thus limit our analysis of YouTube to the comments made on videos. We aim to compare a source that is video-driven to one that is not by uncovering the topical and sentiment leanings in each stream. There are at least two reasons why these might differ: differences in the populations participating and the kinds of interactions offered within each medium. For example, YouTube does not allow users to share links in their comments, whereas Twitter users employ URL shortening services[1] to post links to outside sources.

It is appropriate to compare Twitter and YouTube as both have played an important role in recent political events, as for example in Middle East, United States, and elsewhere [25]. Since our goal is to compare, we focus on a common set of political discussion topics, these include politicians, issues, and political events. Already, some work has been done examining social media dialogue around politicians [22, 84] and events [19]. Unlike previous studies, we also combine politicians and issues to more

[1] such as http://tinyurl.com/

deeply understand the discourse concerning them.

We have several specific goals in this chapter. The first is to re-examine standard approaches and measures seen in the political discourse literature and test them in our two-social media setting. Discussion volume is one that has been used to estimate political favorability of the crowd [84, 57] and sentiment (polarity) counts is another standard one used for similar purposes [19, 26, 57]. Second, we examine in considerable detail the relationship between political stance (or agreement) taken by the author of the text and the sentiment conveyed. We believe that although the two appear to be similar and have been used for similar purposes [19, 57], there are significant differences that should not be ignored. This analysis is another unique contribution of our work. We also examine several stylistic characteristics of the texts such as the presence of humor, sarcasm, and quotations of outside sources. These are recognized as aspects that complicate the analysis of political discourse [23, 79]. In fact some researchers [24, 53], are skeptical about the efficacy of current techniques at tackling the convoluted kinds of rhetoric seen in political discussions. Our focus on stylistic characteristics should provide further insight in this direction. Finally we compare two approaches to sentiment classification: lexicon-driven and data-driven and examine the generalizability of models learned in one source to another. All of these goals are designed to a) understand better the techniques used commonly in analyzing political text and b) understand better the similarities and differences of signals reflected by different social media in the political sphere. To summarize, this project makes these unique contributions:

1. We compare two social media sources – YouTube and Twitter – on a common set of topics. Both in the general SA and in the particular political analysis arena analysis of more than one source is almost never done.
2. We analyze important political topics encompassing current political events, prominent politicians, and popular economic and societal issues.
3. We examine the distinction between agreement (political stance) and sentiment (as typically defined on a positive/negative spectrum) in political writing.
4. We perform a stylistic and linguistic analysis of the two data sources, comparing the levels of sarcasm, humor, quotation, swearing, and linking, as well as word usage.
5. We explore the use of lexicon-driven and data-driven approaches for sentiment classification and model generalization from one source to another.
6. Finally, we contribute an annotated dataset spanning two social media sources and topics including politicians, issues, and events. The annotations are of sentiment and its target, writing styles, and author's position on the topic (agreement).

6.1 Data Collection

Our dataset consists of Twitter posts and YouTube comments on a set of common topics which are of two types. The first is a politician - issue combination. The first two columns of Table 6.1 list the politicians and issues. These yielded a total of 13 × 13 = 169 combination topis. We also studied 9 event topics (listed in

Table 6.1: Politician, issue, and event topics

People	Issues	Events
Barack Obama	Abortion	Occupy Wall Street
Barney Frank	Afghanistan war	Republican Debate
Nancy Pelosi	Debt Ceiling	European austerity measures protests
Elizabeth Warren	Gay Marriage	
Mitt Romney	Health care	Syrian uprising
Herman Cain	Immigration	Berlusconi Resigns
Rick Perry	Iraq war	US Military Presence in Australia
Newt Gingrich	Libya war	
Jon Huntsman	Nuclear Iran	Jobs Bill pass
Rick Santorum	Pro-Choice	Northern Gateway pipeline project
Michele Bachmann	Pro-Life	
Ron Paul	Tax Reform	Personhood amendment in Mississippi
Sarah Palin	Unemployment	

last column of Table 6.1). Each topic yielded a query that was executed both on YouTube and on Twitter collecting YouTube comments and Twitter posts for the period of November 16 to 24, 2011.

We implement a two-step approach to collecting YouTube comments. First, using YouTube Search API we collected the top 50 returned videos. Then, for each video we collected up to 500 most recent comments. Unlike Twitter, YouTube has not been explored widely as a source of textual data. Thus, we test several search strategies in order to collect the best quality data.

6.1.1 Searching YouTube

Because the relevance of YouTube comments we gather depends on the relevance of the videos we gather, we first conducted a small scale study to determine the best approach to retrieve relevant videos. YouTube Search API allows four ways of ranking the results:

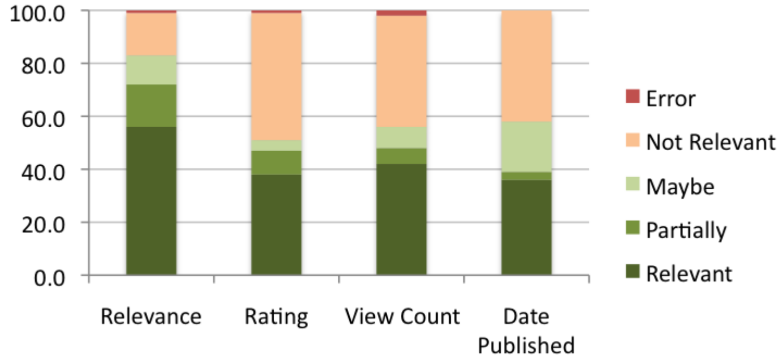


Figure 6.1: Percentage of relevant videos for various rankings

1. Relevance – by relevance of the video, as determined by YouTube search algorithm (likely using video title and other metadata)
2. Rating – by rating of the video, as a proportion of “likes” to “dislikes”
3. Views – by the number of views of the video
4. Date – by the date the video was posted

We evaluated the relevance of the top 20 returned videos for five select queries. They were classified as: *relevant*, *partially*, *maybe*, *not relevant*, and *error* (for example when a video has been removed). Figure 6.1 shows that relevance ranking gave the best performance, with the three other rankings performing similarly. Thus, we use Relevance searches to retrieve videos.

6.2 Discussion Volume

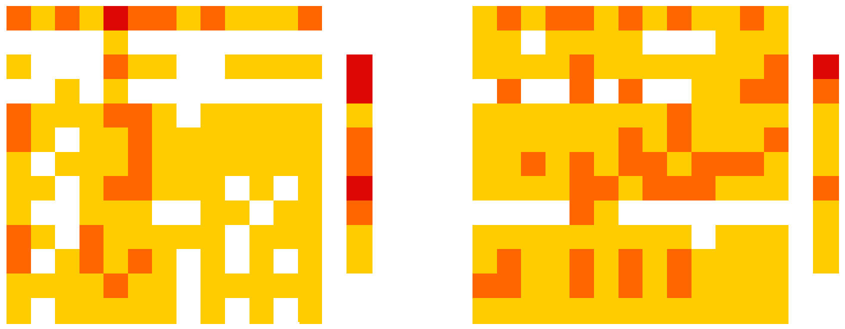
We first examine the volume of documents – tweets (for Twitter) and comments (for YouTube) – returned for each query. We are interested discussion volume as it has been seen to correlate with popularity [84, 57]. The search described above yielded 27,084 tweets and 40,775 YouTube comments with median of 6.5 tweets and 49 YouTube comments per query. Fifty queries run against Twitter and 22 against YouTube returned zero results (with the intersection of 15 queries). For example, five queries with *Elizabeth Warren* did not return any results for either source. On the other hand, there is a substantial amount of chatter around the Republican politicians running for the US Presidential nomination (including *Romney*, *Cain*, *Gingrich*, *Bachmann*, and *Paul*). We also see some issues being discussed along with a particular politician, such pairs include *Gay Marriage* and *Rick Santorum*, *Pro-Life* and *Rick Perry*, and *Unemployment* and *Elizabeth Warren*.

Figure 6.2 shows the distribution of the number of documents that were gathered. The color scheme is as follows: white signifying 0 documents, yellow – under 0.005% of all documents in dataset, orange – under 0.05%, and red – 0.05% of the documents and over. In the square, the rows signify politicians and columns issues ordered from upper left as in Table 6.1. The column to the right shows the nine events. At a glance, we see that the users of these two sites focus on different topics, with 50% of the cells not matching in color.

Yet, because of the difference in data collection approaches for the two streams, it is more informative to examine the volume of topics within the streams. Table 6.2

Table 6.2: Rankings of topics by discussion volume

Politicians			
Twitter		YouTube	
Obama	4917	Obama	6543
Romney	2923	Perry	4923
Gingrich	1553	Elizabeth Warren	4377
Perry	1391	Ron Paul	4039
Bachmann	944	Gingrich	2573
Ron Paul	695	Cain	2407
Santorum	621	Bachmann	2307
Cain	600	Romney	1421
Pelosi	553	Pelosi	734
Palin	72	Jon Huntsman	472
Jon Huntsman	38	Santorum	381
Elizabeth Warren	5	Palin	368
Barney Frank	1	Barney Frank	315
Issues			
Twitter		YouTube	
Immigration	4325	Nuclear Iran	5438
Health care	4064	Iraq war	4435
Abortion	1938	Health care	4179
Unemployment	1017	Tax Reform	3169
Gay Marriage	801	Afghanistan war	2960
Nuclear Iran	771	Unemployment	2254
Debt Ceiling	466	Pro-Life	1979
Pro-Life	344	Libya war	1613
Iraq war	238	Abortion	1421
Libya war	110	Pro-Choice	1138
Afghanistan war	100	Gay Marriage	959
Pro-Choice	72	Immigration	691
Tax Reform	67	Debt Ceiling	624
Events			
Twitter		YouTube	
Occupy Wall Street	5000	Occupy Wall Street	8013
Republican Debate	4494	US Military Presence in Australia	1009
US Military Presence in Australia	1499	Republican Debate	487
Syrian uprising	842	European austerity measures protests	183
Jobs Bill pass	677	Berlusconi Resigns	86
Berlusconi Resigns	206	Personhood amendment in Mississippi	49
Personhood amendment in Mississippi	42	Jobs Bill pass	44
Northern Gateway pipeline project	10	Syrian uprising	22
European austerity measures protests	1	Northern Gateway pipeline project	22



(a) Twitter

(b) YouTube

Figure 6.2: Discussion volume heat maps – rows: politicians, columns: issues; last column: events.

shows the politician, issue, and event queries ranked by the number of documents retrieved, with politician topics aggregated over issues, and issues aggregated over politicians. We compute Spearman's rank correlation coefficient between Twitter and YouTube lists to be 0.566 for politicians, -0.192 for issues, and 0.583 for events. Thus politician and event rankings are more similar across sources than issue lists. And, indeed, looking at the two ranked lists of issues, we see, for example, that the discussion in the YouTube comments about the three mentioned wars (Iraq, Afghanistan, and Libya) is greater than in Twitter. Also, *Debt Ceiling* topic is much more popular in Twitter than *Tax Reform*, and the opposite is true for YouTube. Because the two topics are related, it could be argued that a mere word choice in the discussions of these issues could result in a divergent results. However, we note that *Immigration* is at the top of the list for Twitter and second from the bottom for YouTube, indicating a drastic difference in the discussion volume in the two streams.

Finally, we compare the Republican Party candidates rankings to those produced by the closest Gallup poll[2], one taken around the same time, Nov 13 - 17. Using Spearman's rank correlation coefficient, we get 0.771 for Twitter and 0.314 for YouTube (see Table 6.3), showing Twitter to be better at matching Gallup poll ranking. Considering the joint evidence across the two sources we also rank politicians by the sum of the documents in both sources (column Both), getting a correlation with Gallup poll of 0.428, i.e., not better than Twitter alone. It is interesting to note that this supports the popularity Twitter is gaining for predicting election outcomes. In particular, [84] found that "the mere number of tweets mentioning a political party can be considered a plausible reflection of the vote share and its predictive power even comes close to traditional election polls." However, although fairing better than YouTube, Twitter does not predict the top candidate correctly, as would be a requirement for a successful election predictor.

From these observations, we conclude that the emphasis of the discussions in the two streams is dissimilar, with some topics getting a lot of attention in one, but not in another. This suggests that the user base in Twitter and YouTube either differs widely, or the services are used in a different way to discuss political topics. We further show that neither Twitter nor YouTube predict the republican frontrunner in the US Presidential election, in contradiction to the observation by [84] that discussion volume is enough to predict election polls. This may be due to the fact that discussion volume may also indicate forms of interest other than favorable, for example, in case

[2] http://www.gallup.com/poll/election.aspx

Table 6.3: Rankings of the Republican Presidential nominee hopefuls by Gallup Poll and by discussion volume

Gallup	**Twitter**	**YouTube**	**Both**
Gingrich	Romney	Perry	Perry
Romney	Gingrich	Paul	Paul
Paul	Perry	Gingrich	Romney
Perry	Bachmann	Bachmann	Gingrich
Bachmann	Paul	Romney	Bachmann
Santorum	Santorum	Santorum	Santorum

of a scandal or a tragic event. Also, an especially vocal support base for a politician may inflate the discussion volume, as in the case of Ron Paul, as we discuss later.

6.3 Content Analysis

We further analyze the two streams by labeling a subset of the above topics. We label them for sentiment, agreement, writing style, inclusions of links and also explore vocabulary usage.

6.3.1 Data Selection

We choose topics by taking the intersection between the top seven politician, issue, and event lists (ranked by volume) in both streams which have at least 100 comments/tweets in the dataset. These resulted in the following five politicians: *Obama*, *Perry*, *Gingrich*, *Bachmann*, and *Ron Paul*; three issues: *Health Care*, *Nuclear Iran*, and *Unemployment*; and two events: *Occupy Wall Street* and *US Military Presence in Australia* (excluding *Republican Debate* because of vagueness of the query) (see

Table 6.4: Selected queries

Q#	Query
1	Obama Health care
2	Obama Nuclear Iran
3	Obama Unemployment
4	Perry Health care
5	Perry Nuclear Iran
6	Perry Unemployment
7	Gingrich Health care
8	Gingrich Nuclear Iran
9	Gingrich Unemployment
10	Bachmann Health care
11	Bachmann Nuclear Iran
12	Bachmann Unemployment
13	Ron Paul Health care
14	Ron Paul Nuclear Iran
15	Ron Paul Unemployment
16	Occupy Wall Street
17	US Military Presence in Australia

Table 6.4).

Since we have limited resources for annotation, we first identify items most likely to be relevant by using a search engine to rank the documents. For both streams we index the documents (tweets for Twitter and comments for YouTube) using Lemur Indri[3] resulting in an index for each query, for each stream. We then search the index using the query, resulting in a ranked list of documents. The annotators started from the top of the ranked lists, labeling until 100 relevant documents have been identified, or until at least 20 non-relevant documents have been encountered in a row. The relevant items were then labelled for the various content aspects such as sentiment, agreement etc.

[3] http://www.lemurproject.org/

Table 6.5: Percentage overlap between annotations of select queries

		Relevance	Target	Agreement	Sentiment
YouTube	Obama Nuclear Iran	83.9	70.5	69.5	78.9
	Ron Paul Health care	98.1	94.0	89.0	79.0
	Occupy Wall Street	94.3	–	–	65.0
Twitter	Obama Nuclear Iran	83.8	89.3	92.2	87.4
	Ron Paul Health care	100.0	95.0	87.0	87.0
	Occupy Wall Street	97.1	–	–	72.7
		Sarcasm	Humor	Flaming	Quotation
YouTube	Obama Nuclear Iran	97.3	96.4	94.6	100.0
	Ron Paul Health care	99.0	98.1	97.1	99.0
	Occupy Wall Street	97.1	94.3	90.5	100.0
Twitter	Obama Nuclear Iran	98.2	100.0	99.1	91.0
	Ron Paul Health care	98.0	96.0	99.0	84.3
	Occupy Wall Street	100.0	98.0	99.0	82.4

6.3.2 Labeling

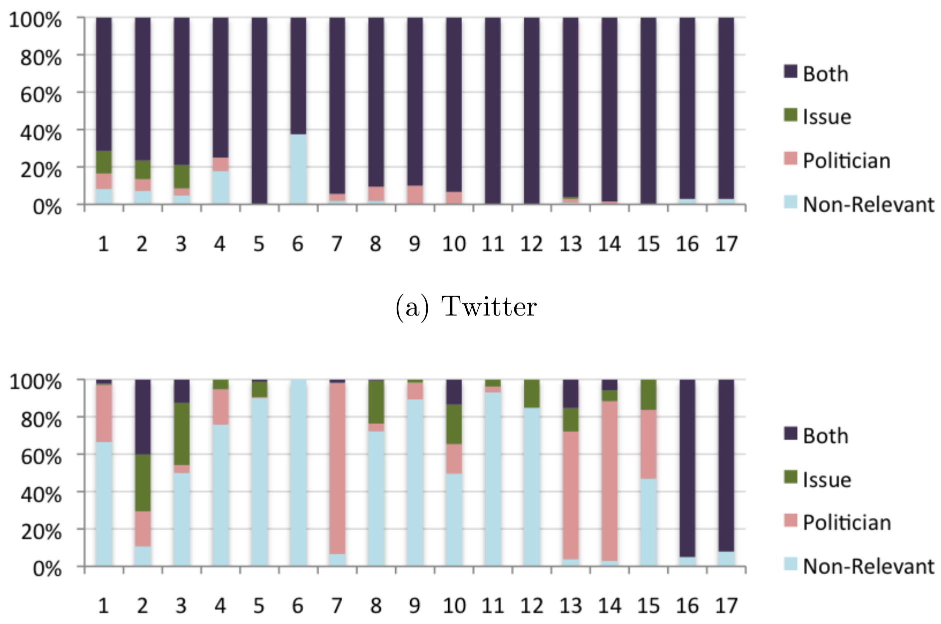
Two annotators (well-versed in political speech) then labeled documents from both streams. After the relevance of the document has been identified, for politician/issue queries annotators marked the target of the writing – politician, issue, or both. Then two labels were determined – whether the author agreed with the politician's stance on the issue (with choices *Agrees*, *Disagrees*, *None*), and the emotional sentiment of the document (*Positive*, *Negative*, *Mixed*, or *None*). Finally annotators noted stylistic features of the text, which included the presence of *Humor*, *Sarcasm*, *Flaming*, and whether the text has a *Quotation*.

Three queries were chosen to be labeled by both annotators in order to compute inter-annotator agreement. The percentages of overlapping label instances for each selection are shown in Table 6.5. We use percentage overlap as inter-annotator agreement measure instead of the standard one like Cohen's Kappa because in many cases the documents belong largely to one class and not the other (that is, the distribution of labels is skewed). Kappa measure (and several others) uses marginal probabilities determined from the labels to estimate agreement achieved by pure chance, making Kappas for skewed data very low or negative, even though there is a large overlap in labels. Indeed, annotators agreed very well on most of the tasks with average *Relevance* overlap of 92.9% and 96% for stylistic labels (rightmost four in the table). *Agreement* and *Sentiment* annotations proved to be more difficult, with Sentiment overlap for *Occupy Wall Street* query being 65% for YouTube. Note that overlap is on average greater for Twitter tasks than for YouTube, perhaps due to YouTube comments being longer and allowing for a more complex expression of sentiment.

6.3.3 Relevance

Figure 6.3 shows the relevance of documents in each query. It shows whether the documents are relevant to the combination of politician and issue ("both") or to just one component. Observe that a vast majority of documents retrieved from Twitter have turned out to be relevant to both the politician and the issue. This accuracy justifies many recent research strategies which use Twitter Search API to collect their datasets [84, 19].

YouTube, on the other hand, provides very few documents about both the issue and politician for first 15 queries, but has a very high accuracy for the two event queries. For politician/issue queries, it does capture conversation about either



(a) Twitter

(b) YouTube

Figure 6.3: Document relevance to selected queries

the politician or the issue. For example, over 70% of the documents about topics 13 and 14 (with Ron Paul) are just about the politician and not the issue; same is not the case, for example, for Bachmann queries (10,11,12). Thus, in the sense that only documents relevant to both the issue and the politician being relevant, YouTube gives us a much worse performance – an average of 15% precision (compared to 89% for Twitter). But if we treat discussion about issue only or politician only as also relevant, we get precision of 95% for Twitter and 49% for YouTube. Overall, retrieval of relevant comments for YouTube was a harder task.

Table 6.6: Overall sentiment in each stream (percentages)

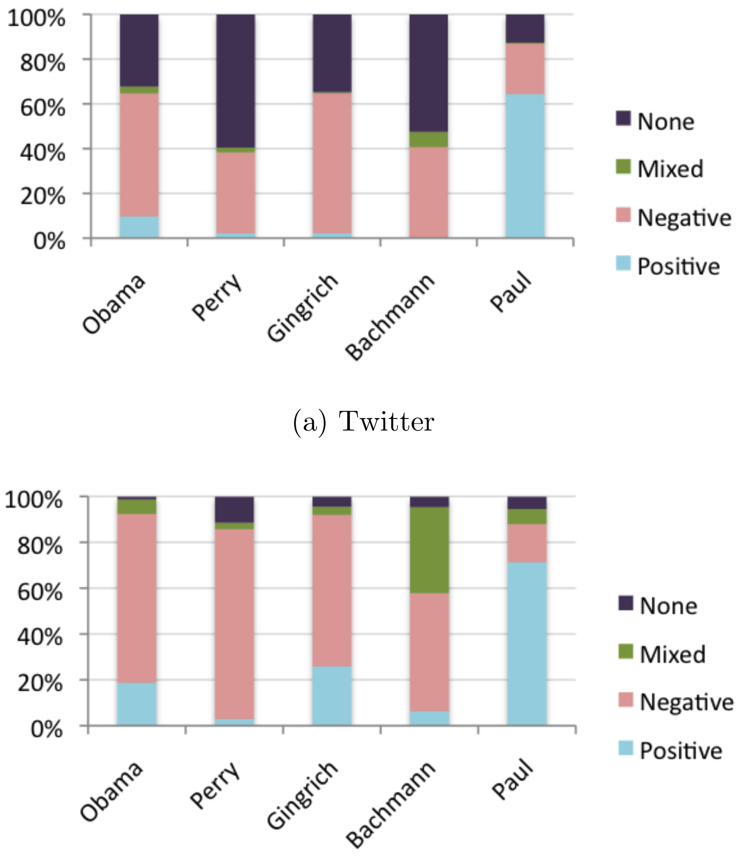
	Pos	Neg	Mix	None
Twitter	17.5	40.6	1.8	40.1
YouTube	27.7	59.0	6.6	6.7

6.3.4 Sentiment

Table 6.6 shows sentiment expressed by the documents about politicians, issues, or their combination. The two streams differ drastically in the number of documents showing no sentiment (column None). Otherwise, the proportion of positive to negative sentiment is similar between the two streams, favoring negative about 2 to 1. So there is consistency between the two streams in this regard. As we later show, negative sentiment dominates all discussions, both those about liberal or conservative politicians, and along all of the issues. Thus, it may be the case that the default tone of any political discussion is negative irrespective of medium.

We further examine the sentiment expressed about the issues and politicians by aggregating appropriately. Figure 6.4 shows the sentiment of documents talking about politicians. The politician getting the most positive sentiment in both streams and by a large margin of difference from the next politician is Ron Paul. This is consistent with the fact that he is known for his active and young base[4]. But there are notable differences between the streams. For example, YouTube shows over 20% positive

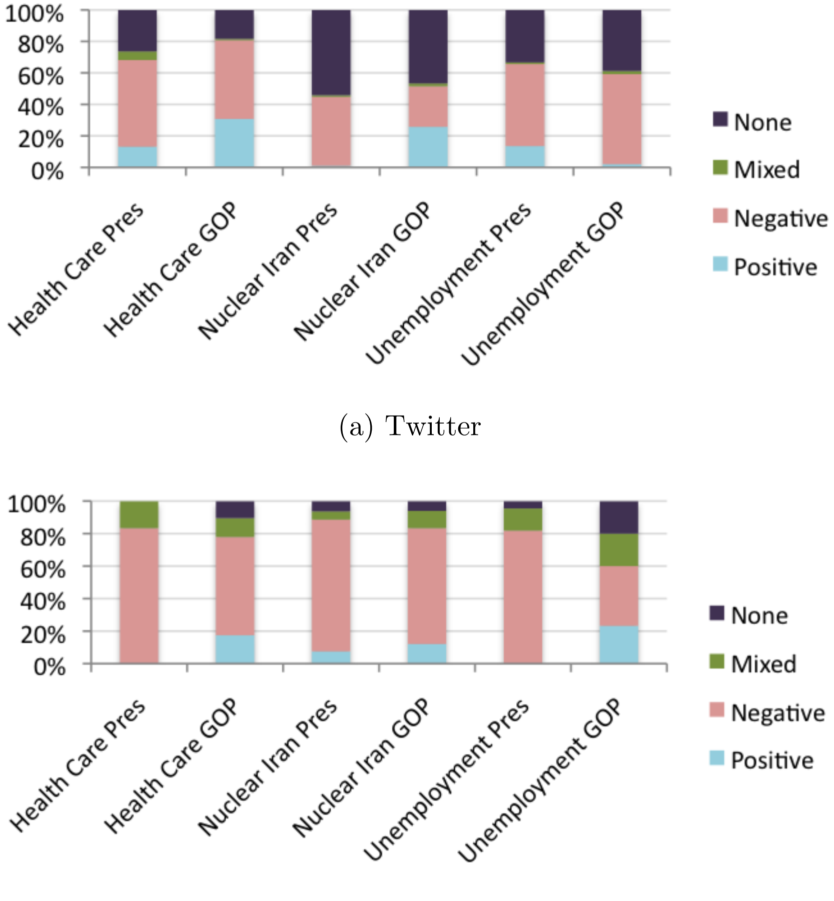
[4] http://www.huffingtonpost.com/2012/01/12/ron-paul-young-voters_n_1202616.html



(a) Twitter

(b) YouTube

Figure 6.4: Sentiment summaries of politicians



(a) Twitter

(b) YouTube

Figure 6.5: Sentiment summaries of issues

sentiments about Newt Gingrich, but his support is near zero in Twitter. And, over a third of the YouTube comments about Bachmann express mixed sentiment (compared to 4% for Twitter), showing that in her case the discussion on YouTube can be more complex. Thus the two sources express different sentiment signals for these politicians. This is also observed when we compare the negative sentiments expressed. Perry takes the lead in YouTube for this but not so in Twitter.

We take a different approach to comparing sentiment across issues. Since the political party may take opposing positions, we divide the data into 2 groups: President (who is considered liberal or centrist) and GOP (who are considered conservative) – all other politicians are in GOP. The summaries are shown in Figure 6.5.

Note the overwhelming negative sentiment in YouTube (on average 81% for Pres and 60% for GOP), which is less so in Twitter (on average 50% for issues relative to Pres. and 42% for issues relative to GOP). Furthermore, there is more positive sentiment for GOP side of the issues, except for Unemployment, where in Twitter President gets more positive signals, whereas in YouTube GOP's stance is favored more. This shows that the two media differ in the sentiment signals, and also that some issues may polarize people differently on different social media.

Note, also, that most of these documents contain very few mixed sentiment documents. This is likely because of the limited space allowed for these writings (140 for Twitter and 500 for YouTube), which discourages that kind of discourse. The average length of YouTube comment is almost twice as much as that of a Tweet (220

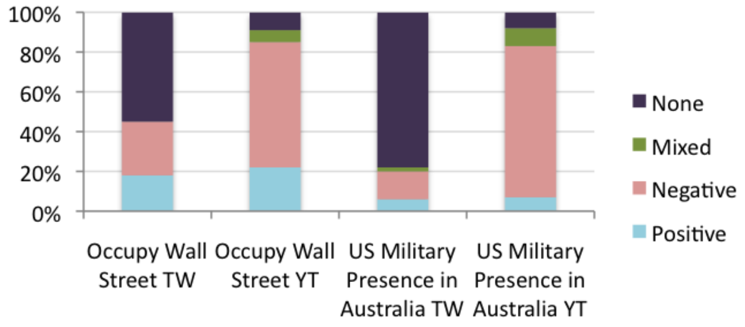


Figure 6.6: Sentiment distribution of events

Table 6.7: Republican Presidential nominee hopefuls rankings (by sentiment)

Gallup	**Twitter**	**YouTube**
Gingrich	Paul	Paul
Paul	Perry	Gingrich
Perry	Bachmann	Bachmann
Bachmann	Gingrich	Perry

compared to 122 characters), and, YouTube has 6.6% mixed documents compared to 1.8% in Twitter. This suggests the obvious that longer space allowance might foster a discussion that considers both sides of the issue.

Finally, we examine sentiment expressed about the two events – Occupy Wall Street and US Military Presence in Australia (see Figure 6.6). For both queries, YouTube comments express much more sentiment than their Twitter counterparts, and in both cases, more negative sentiment (though positive sentiment remains similar). Similar to other queries, YouTube has a few more mixed-sentiment documents.

Although both streams favor negative to positive sentiment (roughly 2 to 1), our analysis reveals differences in sentiment of the discussion between the two sources. For example, YouTube shows more support for GOP stance on Unemployment, whereas Twitter discussion favors the President. YouTube discussion of the two selected events also shows a stronger bias toward a negative sentiment. These sentiments, however, do not reflect the general sentiment as may be determined using traditional polling methods. As for discussion volume, we compared the Gallup poll GOP politician rankings to the rankings of our select GOP politicians (ranked using sentiment), and found that neither predicted the frontrunner, and both overestimated the popularity of Mr. Ron Paul (see Table 6.7). Comparing these ranks to Gallup poll ranking, we find Spearman's rank correlation coefficient -0.199 for Twitter and 0.60 for YouTube. The interesting point here is that with discussion volume Twitter did much better than YouTube (0.771 correlation compared to 0.314), here the performance with sentiment is reversed.

6.3.5 Agreement versus Sentiment

Besides discussion volume and sentiment, another aspect being examined is that of stance taken by the text w.r.t the politician or issue of interest. We use the term 'agreement' for this as it more clearly indicates whether the stance taken by the author of the text agrees or not with the stance of the politician or issue (i.e., the topic). We see for example, sentiment and agreement used for similar purposes [19, 43, 57, 78]. The more positive the sentiments expressed the greater the support inferred and counts of agreement also indicate the overall level of support. Of the two, the problem with sentiment is that it is not enough to identify the sentiment conveyed in the text. It is also important to identify the target of the sentiment and make sure that it is "on topic". This point about the importance of the target was usually understood in research involving sentiment about products or movies [45], but it appears to have become somewhat lost when it comes to analysis of political discussions. We see for instance, papers where counts based on a sentiment lexicon are used to estimate sentiment and thereon to estimate support [57]. To further understand this aspect we compare these two approaches and determine the extent to which they run parallel to each other.

First we look at the number of documents that express sentiment but this sentiment is not on topic. For example, consider the topic of *Newt Gingrich* and the document *"White House is full of liars and old scoundrels, makes me sick! Vote-NEWT12"*. The document conveys negative overall sentiment (given words such as *liars*, *scoundrels*, *sick*) but this is not directed to towards the topic, and hence we regard this sentiment as not being on topic. Had the topic been *the White House* the sentiment would have been on target and negative. We find 56 tweets (7.3% of total) and 176 YouTube comments (24.3% of total) to have sentiment that is off topic. 7% may seem to be a small value however, this indicates that the margin of error is somewhere between 7 and 14% for Twitter when comparing sentiment across two topics. The range is much higher for YouTube.

We now look at the relationship between agreement and sentiment. We define two categories. The first category consists of combinations of agreement and sentiment that are synchronized. These include "agrees + on topic positive sentiment", "disagrees + on topic negative sentiment" and also "neutral stance + on topic neutral sentiment". The second category includes all other combinations of agreement and sentiment, as for example, "neither agrees nor disagrees + on topic negative sentiment" and "agrees + off topic negative sentiment". Note in all of these, sentiment refers to the dominant sentiment expressed in the document. We note that for YouTube only 66% of the documents (480/725) fall into the desired category. The remaining 34% are noisy in this regard. In Twitter, 89% of documents are in the desired category while the remaining 11% are in the noisy category. Thus we see that there is non trivial noise present and again more so in YouTube than in Twitter. We present some examples of documents in the noisy category.

Topic: Bachmann on Health Care

"when politicians f— with our money thats a problem when they f— with our health .....thats a whole different level"

Agrees/Negative: The document is in agreement with small government approach to health care of the politician, but overall is negative toward government intervention.

Topic: Ron Paul on Nuclear Iran

"Yup, & a sanctioned Nuclear Islamic Iran RT @JamesWolcott:At least

with Ron Paul you know you're not going to get a sob story. #tff11"

Neutral/Positive: The author's agreement with Paul's stance on Iran is unclear, but the later part of the tweet is positive.

One of the most notable noisy category was "neither agrees nor disagrees + negative sentiment" – a negative banter which is relevant to the topic, but does not express a definitive stance – comprising of 12.3% of the YouTube comments. For example,

Topic: Ron Paul

"If Ron Paul becomes President, he WILL be assassinated."

Neutral/Negative: This YouTube comment is pessimistic about Ron Paul's safety as a President, but does not state explicitly an agreement or disagreement with Mr. Paul's point of view.

Within the limits of this study, it is 89% safe to assume that agreement is the same as sentiment in Twitter. It is only the case in 66% of the time in YouTube comments. In longer and more complex writings, this distinction may be even more pronounced. Therefore, the definition of "sentiment" in political discourse should be delineated clearly to distinguish between political opinions and emotional states, lest one is misinterpreted as another and inaccurate conclusions are made. These results also indicate that simple lexicon based classification of sentiment is likely to be of limited value in political discourse.

6.4 Language and Style

6.4.1 Style

Table 6.8 shows statistics on various stylistic features of the text. Twitter has many more quotations across all sentiment classes than YouTube. Negative documents have higher chance of being sarcastic, but this is not a very dominant trait in either dataset. Flaming (using inflammatory language) happens more in the Negative ones, though in YouTube it also occurs in other sentiment classes.

Table 6.8: Stylistic features (% of documents in sentiment class)

	Twitter				YouTube			
	Pos	Neg	Mix	Non	Pos	Neg	Mix	Non
sarcasm	1.2	7.1	0.0	0.3	1.6	5.0	1.6	1.6
humor	0.6	1.5	5.9	0.8	2.0	3.5	6.6	9.7
flaming	0.6	2.3	0.0	0.0	2.7	10.5	3.3	3.2
has quote	55.3	44.4	29.4	50.4	0.0	0.6	3.3	1.6

It has been observed by [79, 5] that people use sarcasm and humor to make their point in ideological arguments, making them more challenging to analyze. However, we show that, even though these are present, they are not dominant in our dataset.

6.4.2 Language

We examine the text itself by building language models for each query's documents. Using a lingpipe[5] tokenizer we extracted the 1, 2, and 3-grams.

Table 6.9: Language and vocabulary statistics

	# tokens	# unique words	# docs	unique wds/doc	unique wds/toks	# dups
Twitter	60069	31107	970	32.07	0.517	368
YouTube	113526	79493	923	86.12	0.700	4

Table 6.9 shows that compared to the amount of text (# of tokens), YouTube contains more unique words than Twitter: 0.70 unique words per token in YouTube, compared to 0.52 unique words in Twitter. This redundancy may be due to Twitter containing 37.9% near-duplicate tweets (by means of re-tweeting). We also looked at the use of links in the text, and saw a very different behavior between the two streams. YouTube does not allow to share URLs, but it is possible to do so by inserting characters into the URL or leaving some of it out. We found 1.4% of YouTube comments 79.7% of tweets had URLs, confirming that the nature of discussion on these two media is different – one is meant largely to share opinion (YouTube), the other also information (Twitter). These informational tweets were more often coded as having sentiment: 59.4% Pos/Neg/Mix than 40.6% None. Interestingly, the same proportion

[5] http://alias-i.com/lingpipe/

applies to tweets without URLs: 61.4% Pos/Neg/Mix to 38.6% None, meaning that URL is not a distinguishing feature of sentiment-laden speech on Twitter.

6.5 Classification

6.5.1 Lexicon-Driven Classifiers

Using our data, we examine a popular sentiment classification approach which uses standard lexicons such as in [57, 84, 7, 26]. Our concern is that general-purpose lexicons used in these studies are not suitable for political speech, yet they are often used without evaluation. Moreover, as we have just shown the sentiment can be off topic. We test a sentiment classifier using SentiWordNet [21] – a collection of 52,902 words from the WordNet database automatically annotated with a positive and negative score (both ranging from 0 to 1).

Table 6.10: SentiWordNet sentiment classifier performance

		Positive			Negative		
	Acc	Prec	Rec	F	Prec	Rec	F
Twitter	0.624	0.256	0.129	0.172	0.690	0.838	0.757
YouTube	0.591	0.321	0.277	0.297	0.691	0.734	0.712
TW majority	0.699	0.000	0.000	0.000	0.699	1.000	0.823
YT majority	0.688	0.000	0.000	0.000	0.688	1.000	0.815

Table 6.10 shows overall accuracy and precision, recall, and F-score for both polarities. For comparison, we also show the majority vote baselines for each stream

(predicting majority class for all documents). Judging by accuracy, in both cases the baselines perform better than our classifier. The problem is especially with the positive class, since so many words deemed negative appear in positive documents. A lexicon tailored to political speech may perform better, and we leave the development of such tools for future research.

6.5.2 Data-Driven Classifiers

An alternative method to using pre-defined lexicons is building classification models using the features extracted from text. We examine this approach by building classifiers for both streams using computational linguistics toolkit Lingpipe[6]. Using its tokenizer, we extract 1-, 2-, and 3-grams (sets of consecutive words) and build a logistic regression classifier to label the sentiment of each document. We approach the classification task as in the previous chapter by distinguishing between two tasks:

- *POS*: target class includes Positive and Mixed sentiment documents (as opposed to Negative and None)
- *NEG*: target class includes Negative and Mixed sentiment documents (as opposed to Positive and None)

We test the classifiers using leave-one-out strategy, in which the classifier is trained on all but one document and tested on that document, and this is done for all documents in the dataset.

[6] http://alias-i.com/lingpipe/

Within stream classification. Table 6.11 shows the performance of the Lingpipe classifiers in each stream and for each task. The performance is much better than that of lexicon-driven classifier above. The performance is especially impressive for the positive classifiers, which are identifying a minority class. Though the task is especially challenging in YouTube, with F-measure at 0.577.

Table 6.11: Lingpipe sentiment classifier performance within each stream

Stream	Task	Accuracy	Precision	Recall	F-measure
Twitter	POS	0.912	0.836	0.679	0.749
Twitter	NEG	0.822	0.816	0.747	0.780
YouTube	POS	0.750	0.689	0.497	0.577
YouTube	NEG	0.758	0.765	0.911	0.831

Within topic classification. It may be the case that conversation is so peculiar for each politician/issue combination that aggregating over all of the data hurts performance. Instead, we build classifiers for each topic individually. We choose topics having at least 40 relevant documents in both streams, resulting in five topics shown in Table 6.12. The performance varies greatly, with some measures being affected by the smallness of the positive class, where the POS F-measures are 0.000. In other cases, especially for the negative class, the F-measures get as high as 0.938 for Twitter and 0.942 for YouTube. However, on average topic-specific classifiers perform worse than the aggregate one above with average F-measures of 0.392 (POS) and 0.720 (NEG) for Twitter and 0.430 (POS) and 0.718 (NEG) for YouTube.

Table 6.12: Lingpipe sentiment classifier performance within each stream and topic

Twitter		Accuracy	Precision	Recall	F-measure
Obama Health care	POS	0.870	0.727	0.444	0.552
Obama Health care	NEG	0.730	0.743	0.873	0.803
Obama Nuclear Iran	POS	0.942	0.000	0.000	0.000
Obama Nuclear Iran	NEG	0.854	0.837	0.854	0.845
Ron Paul Health care	POS	0.870	0.901	0.914	0.908
Ron Paul Health care	NEG	0.900	0.700	0.500	0.583
Ron Paul Nuclear Iran	POS	0.862	0.864	0.927	0.894
Ron Paul Nuclear Iran	NEG	0.877	0.833	0.750	0.789
Gingrich Health care	POS	0.971	0.000	0.000	0.000
Gingrich Health care	NEG	0.886	0.909	0.968	0.938
Gingrich Nuclear Iran	POS	0.962	0.000	0.000	0.000
Gingrich Nuclear Iran	NEG	0.865	0.500	0.286	0.364
YouTube		Accuracy	Precision	Recall	F-measure
Obama Health care	POS	0.662	0.625	0.192	0.294
Obama Health care	NEG	0.718	0.714	0.957	0.818
Obama Nuclear Iran	POS	0.920	1.000	0.200	0.333
Obama Nuclear Iran	NEG	0.890	0.890	1.000	0.942
Ron Paul Health care	POS	0.680	0.731	0.778	0.754
Ron Paul Health care	NEG	0.730	0.606	0.588	0.597
Ron Paul Nuclear Iran	POS	0.740	0.771	0.949	0.851
Ron Paul Nuclear Iran	NEG	0.780	0.500	0.136	0.214
Gingrich Health care	POS	0.740	0.778	0.226	0.350
Gingrich Health care	NEG	0.730	0.728	0.971	0.832
Gingrich Nuclear Iran	POS	0.829	0.000	0.000	0.000
Gingrich Nuclear Iran	NEG	0.829	0.829	1.000	0.907

Cross-stream classification. Next, we assess the extent to which sentiment models learned from one stream can be used to classify documents from another. The results are shown in Table 6.13. In general, performance is quite worse than that of native classifiers (those trained on the same source as the testing data). The only case of close performance is the NEG classifier trained on Twitter and tested on YouTube with F-measure of 0.719 (compared to 0.831 of native classifier). These results are unlike those found in our experiments in the previous chapter, where we show that classifiers trained on reviews or Twitter may perform as well as native classifiers when adapted to other social media sources. However, the topics we explored in that study do not include political discourse.

Table 6.13: Lingpipe sentiment classifier performance across streams

		Accuracy	Precision	Recall	F-measure
Twitter to YouTube	POS	0.694	0.736	0.168	0.273
	NEG	0.629	0.714	0.723	0.719
YouTube to Twitter	POS	0.694	0.343	0.642	0.447
	NEG	0.553	0.458	0.307	0.367

We further examine the difficulty of adapting political sentiment classifiers. On average, when the native and foreign classifiers agree, it is 83% likely that they are correct. When the two disagree on their labels, we have two cases – when the native classifier is correct, or when the foreign one is. Out of such disagreements, it is much more likely that the native classifier is correct (74%) than otherwise. Still, in nearly 8% of all experiments foreign classifiers get the class right and the native do not.

These cases may be explained by vocabulary mismatch. The words used in the document of interest may be unusual for the native stream, but a classifier trained on a foreign stream may have encountered them. Thus, we examine classifier disagreements in terms of vocabulary match, as shown in Table 6.14. Here, we show the average percentage overlap of the document's vocabulary with that of the Twitter model that is classifying it and YouTube model. We also show the overlap with top 1000 terms in each model (determined using SVMlight[7] weight features) – being the most important terms for distinguishing between the classes.

Table 6.14: Percent vocabulary match in cases where native and foreign classifiers disagree

		in Twitter,	top 1000	in YouTube,	top 1000
Twitter to YouTube	native correct	29.3	11.0	83.1	15.3
	foreign correct	30.3	10.8	99.0	18.9
YouTube to Twitter	native correct	93.9	39.3	29.1	10.3
	foreign correct	98.8	21.0	34.7	12.6

As expected, the texts match very well with the models from the native stream. We do see that in some cases a foreign classifier works better because of a better match for its top 1000 features. For example, when adapting YouTube classifier to Twitter, we see a higher match in top 1000 terms in YouTube in cases where YouTube is correct (12.6%) as opposed to when Twitter is (10.3%). Curiously, same is not true when adapting Twitter to YouTube, so even without superior vocabulary match, Twitter classifier can outperform native YouTube one. Below are some examples of cases in

[7] http://svmlight.joachims.org/

which classifiers disagree:

- (Neg.) "*@ObamaNews Obama plz don't be a lapdog for WS. Iran has no "nuclear program." Theres no evidence of one in your press release. #nowaroniran*"

  Adapting YouTube to Twitter, YouTube got the class right. The top 1000 features in Twitter model does not have features "has", "no", or "nuclear", whereas YouTube model does, capturing conversation in which authors say Iran has no nuclear program (which is much more popular stance in YouTube comments than on Twitter).

- (Neg.) "*American presence = TAKEOVER = NWO*"

  Adapting Twitter to YouTube, Twitter got the class right. YouTube 1000 list lacked word "takeover" whereas Twitter list had this rather negative word.

- (Pos.) "*Congrats Obama...something is better than nothing so i'll take it for now http://t.co/5voDHYv5*"

  Adapting YouTube to Twitter, YouTube got the class right. An example of conversational tone which is rare in Twitter training data but is captured in YouTube comments.

6.6 Discussion

The analysis above shows some striking differences between the two streams. Twitter is easier to search, but it provides a redundant set of documents, 40% of which do not contain sentiment. The sentiment that is present is highly polarized with very few mixed sentiment documents. Political discussion on Twitter is overwhelmingly driven by outside sources, with nearly 80% of tweets containing a URL (compared to 13% for general discussion, determined using a 5,6 million general Twitter sample). On the other hand, YouTube is sentiment laden with 93.3% of collected comments contained sentiment. Also, we retrieved a more diverse discussion of the issues and politicians. Overall there were 15 topics for which we could not find any documents in either source and the most discussed topics were about *Occupy Wall Street* movement, *Heath Care*, and *Barack Obama*.

We find the overall sentiment leaning of the two streams to be negative (roughly 2 negative to 1 positive document). The majority of the positive documents, incidentally, came from the supporters of US Congressman Ron Paul. So dominant was this sentiment, that Mr. Paul was at the top of the sentiment rankings in both streams. For two out of the three issues we examined – *Health Care* and *Nuclear Iran* – we saw more positive sentiment about the republican politicians' stance instead of those of the US President. It may be the case that the race for Republican party Presidential nomination has stirred some discussion about their political stances. However, Twitter showed more positive sentiment about the President's stance on the *Unemployment* issue than for Republicans'. This phenomena may also be explained by its contemporary political climate – around the same time President Obama has been active in promoting his jobs plan[8], which may have reflected in greater discussion of

[8] http://www.nytimes.com/2011/11/08/us/politics/senate-acts-on-two-pieces-of-

the topic in positive light. These demonstrate are ability to compare topics across streams and hone in on differences (or similarities) of response across the two social media.

Neither sentiment nor volume of the discussion, reflects the general sentiment as determined using traditional polling methods. Compared to the Gallup poll taken around the same time, neither approach was able to pick out the GOP Presidential frontrunner, putting in question the connection between these social media features and overall political sentiment, as, for example, was observed in [84]. It may be the case, then, that because political discourse in social media is so real-time, it is best used for tracking sentiment on latest events. For instance, it would be interesting to examine the extent to which Twitter focuses on latest events as opposed to general ideological issues.

Our findings about the distinction between *agreement with political stance* and *emotional sentiment* offer insights into the subtleties of political sentiment analysis. We find that in YouTube, 24% of the time when sentiment is expressed it is not on topic. Moreover, agreement and sentiment matched only in 66% of the comments. Though the situation is better with Twitter, it is still present. Thus relying on sentiment expressed in a text alone to determine support for an political issue or personality is a limited strategy (as for example, in [19, 57]). Instead, emotional sentiment detection should be only a part of the analysis, alongside components to examine the target of the opinion.

obamas-jobs-plan.html

Furthermore, we perform a series of sentiment classification experiments. Our preliminary testing of a lexicon-driven sentiment classifier shows that such standard analysis (which has been used, for example, in [7, 57, 84]) is not well suited for sentiment analysis of political discourse. Words deemed negative by the lexicon dominated both positive and negative documents in both streams, biasing the classifier toward the negative class and providing poor performance. It is clearly risky to use lexicons without prior evaluation on political texts (such as in [7, 84]). Classifiers trained directly on the data performed much better, as we show using a logistic regression classifier. Adapting classifiers from one source to another (both from Twitter to YouTube and from YouTube to Twitter) proved to be difficult, with resulting performance much worse than that of the native classifiers. This, again, shows the peculiarity of political sentiment discourse and the difficulty of automated classification. However, we do show that in some instances classifiers trained on outside data would outperform native classifiers by modeling words and stylistic features which are uncommon in the native stream.

6.7 Conclusion

6.7.1 Summary of Findings

In this study we compared YouTube comments and Twitter posts on a set of topics in the domain of politics. Our study indicates several significant differences. The volume of discussion, the amount of sentiment expressed, the nature of agreement vis-à-vis sentiment expressions, all of these show differences across media. Neither medium matches well with Gallup polls. With volume of discussion Twitter appears to have the edge, while with YouTube sentiment does better. A key conclusion is that choice of social medium to analyze determines the results we get. Although we obtain some signals from each that parallel the political world, overall the results obtained across the two media are not consistent. We also studied the relationship between agreement and sentiment and show, for example, that with YouTube we face a greater risk in terms of lack of congruence between the two.

Finally, our test of a standard classifier seen in the political discourse literature indicates risks as well. Using general-purpose lexicons for sentiment classification, as is popular in the literature, results in poor performance. Instead, training classifiers on annotated data proves to be a better choice. However, the choice of training data may affect the resulting performance. Unlike in the previous study in which consumer product topics such as movies and cell phones were used, we show that models trained on a foreign social media source do not perform well compared to those trained on the target data source. We do, however, find that in some cases foreign classifiers may be useful, especially in classifying text with unusual word choice or style, and we leave such exploration to future work.

CHAPTER 7
TRACKING POLITICAL SENTIMENT IN TWITTER

To further understand the nature of sentiment in political discourse in social media, we examine the discussion surrounding the 2012 GOP Presidential candidate selection process. Throughout 2011 and 2012, the US Republican party chooses a nominee for the 2012 Presidential election. This process is highly public, and includes many television appearances and debates.

Analysis of such data is important, considering the attention social media chatter has been getting in the news. Following Barack Obama's 2008 Presidential campaign, the world saw the "crowning of the Internet as the king of all political media" [85]. Online activity indicators such as number of fans on Facebook, followers on Twitter, and likes on YouTube have been seen as indicators of a galvanized base, which ultimately contributed to Obama's victory [55]. Since then, not only has traditional media started paying more attention to political discussions on social media, but several research papers have been published claiming a connection between social media and public polls and even election outcomes. For example, Tumasjan et al. [84] examine Twitter messages about the 2009 German federal election and find that the mere number of messages reflects the election result and even comes close to traditional election polls, concluding that Twitter can be considered a valid indicator of political opinion. Saez-Trumper et al. [72] further improve on this approach by considering only the unique authors, removing the influence power-users would have on perceived conversation volume.

However, some studies show little correlation between the sentiment found on Twitter to that of general public as measured using standard polling techniques. For example, O'Connor et al [57] find no correlation between the 2008 election polls and the support seen in Twitter messages for Obama. They further find that sentiment for McCain (Obama's rival) and Obama slightly correlate (instead of being inversely related). They do find the discussion volume for Obama to have a correlation to the polls, postulating that simple attention may be related with popularity, at least for Obama. Some researchers have came out cautioning against treating social media as a "black box" and letting wishful thinking cloud the analysis of sentiment in political sphere [24]. Metaxes et al. [53], for example, find that electoral predictions using various previously published methods on Twitter data is no better than chance. Among these techniques are discussion volume, lexicon-driven sentiment classification, and user-specific political leaning estimation. Still, some researchers continue to use lexicon-driven classifiers without evaluating their performance [7, 26, 84]. Thus, instead of using a lexicon-driven system, in this study we implement and test a data-driven political sentiment classifier, showing that an effective alternative method is possible. Recently, similar systems employing data mining techniques have been used to identify stances in ideological debates [78] and predictive opinions [37]. Instead, we provide an evaluation of a highly optimized multi-stage approach designed for general-purpose political sentiment classification.

Political sentiment classification is further complicated by differing user behaviors, as discovered by Mustafaraj et al. [55]. Using the 2010 Massachusetts special election for the US Senate, they distinguish between two groups of users: a *silent majority* of people who post very few messages and a *vocal minority* who aim to be read and their messages to be propagated through the Twittersphere. The vocal minority users link more to outside content, use more hashtags, and retweet more. They conclude that aggregating the tweets of both groups may produce a mistaken view of the discussion. Contributing to this stratified group analysis, we also examine each group's writing style, including sarcasm, humor, swearing, and quoting.

In short, this project contributes the following to the political analysis of social media:

1. We build and optimize a multi-stage data-driven sentiment classifier.
2. We analyze sentiment expression in a large sample of Twitter messages, and show the differences between groups of users varying in posting frequency.
3. We perform sentiment tracking experiments in which we compare the sentiment found before and after 19 debates to public opinion polls.
4. We contribute an annotated dataset spanning the second half of 2011 and seven popular Republican Presidential nominee candidates, totaling 6,400 documents annotated for relevance, sentiment about the politician, sentiment intensity, and various stylistic measures.

7.1 Republican Candidate Twitter Data

The data set has been collected by the University of Iowa Political Science professor Dr. Bob Boynton[1] by querying Twitter using Twitter Search API, collecting tweets mentioning various politicians. The collected tweets span a year, over the period of January 1, 2011 to January 11, 2012 and include tweets about the politicians listed in Table 7.1. These are some of the Republican politicians who joined the race for Republican nomination for US Presidential Election of 2012. Some of these joined later in the year, and thus tweets about them do not span the full year. Figure 7.1 shows the discussion volume (in number of tweets mentioning the politician in a week). Notice that the discussion becomes more lively towards the end of the year. Guided by these trends, we select a time span in which to sample the data for each politician – choosing months in which sufficient posting activity is seen. The rightmost two columns of Table 7.1 show the time spans and the number of tweets sampled from that time span. The sampling was done in a random uniform fashion within each month.

The subset, totaling in 6,400 tweets was annotated by a group of political science students as a part of class project. The web interface is shown in Figure 7.2, and the full guidelines can be found in Appendix D. The annotators were given the name of the target politician and a set of tweets. For each tweet, she decided whether the tweet was about the politician by making a *Relevance* judgment. If the tweet was relevant, he/she would decide on whether the tweet was *For* or *Against* the politician,

[1] http://www.boyntons.us/

Table 7.1: Republican Presidential nomination race dataset statistics

	Full Dataset		**Annotated Subset**	
Politician	**Time Span**	**# Tweets**	**Time Span**	**# Tweets**
Michelle Bachmann	1/13/11 - 1/11/12	2,006,034	6/1/11 - 1/1/12 (excluding Oct.)	1,400
Newt Gingrich	1/1/11 - 1/11/12	1,725,271	11/1/11 - 1/1/12	600
Herman Cain	5/26/11 - 1/11/12	1,514,739	9/1/11 - 1/1/12	1,000
Rick Perry	5/27/11 - 1/11/12	1,641,646	7/1/11 - 1/1/12	1,400
Mitt Romney	1/4/11 - 1/11/12	3,170,260	10/1/11 - 1/1/12	800
Ron Paul	1/5/11 - 1/11/12	2,342,392	10/1/11 - 1/1/12	800
Rick Santorum	1/2/11 - 1/11/12	1,125,602	12/1/11 - 1/1/12	400

had *Mixed* opinion, or was *Neutral*. We also allowed for a *Can't Tell* option. Notice that, in the light of our findings in the previous project, here instead of using the standard definition of sentiment as positive or negative, we define it specifically as an opinion about the politician. Furthermore, if the tweet was *For* or *Against* the politician, the annotator needed to select the *Intensity* of the opinion. Finally, several stylistic features of the text were collected: whether tweet contained *Sarcasm*, *Humor*, *Swearing*, or a *Quote*. Some of the tweets were annotated by several (maximum of three) annotators, and majority vote or third annotation broke ties in the cases of disagreement.

Table 7.2 shows annotator agreement as percentage overlap of the labels. The most difficult tasks proved to be *Sentiment* and *Intensity*. Because these are not binary tasks (for sentiment, for example, there are five classes), these numbers are reasonable. Thus, we look at the sentiment data in steps: first we determine agreement in subjectivity (distinction between $\{$*For*, *Against*, *Mixed*$\}$ and $\{$*Neutral*$\}$), then

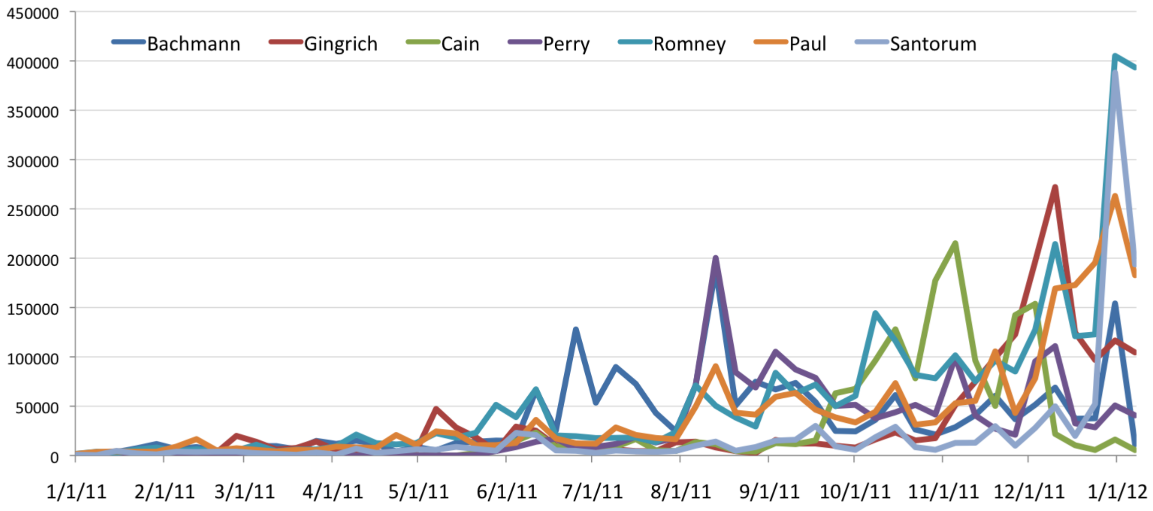
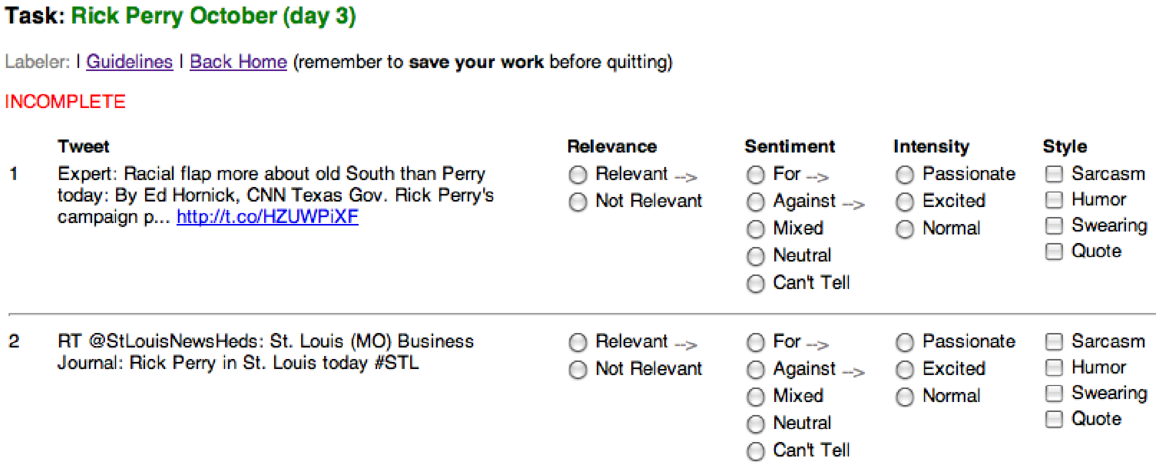


Figure 7.1: Tweet volume for individual politicians in full dataset

in polarity (*For* versus *Against*). Subjectivity proves to be a harder task than polarity. That is, once it is know that the tweet is subjective, it becomes easier to gauge polarity. The labeling interface also allowed annotators to resolve some of their disagreements. Out of these, 19.0% were about *Relevance*, 73.1% about *Subjectivity* and only 7.9% about *Polarity*.

In all, these figures are slightly less than the annotator agreement seen in YouTube and Twitter comparison in previous chapter where *Agreement* percentage overlap ranged from 69.5% to 92.2%. However, these figures put an upper bound to the performance we would expect from our automated classification algorithms.



Task: Rick Perry October (day 3)

Labeler: | Guidelines | Back Home (remember to **save your work** before quitting)

INCOMPLETE

	Tweet	Relevance	Sentiment	Intensity	Style
1	Expert: Racial flap more about old South than Perry today: By Ed Hornick, CNN Texas Gov. Rick Perry's campaign p... http://t.co/HZUWPiXF	○ Relevant --> ○ Not Relevant	○ For --> ○ Against --> ○ Mixed ○ Neutral ○ Can't Tell	○ Passionate ○ Excited ○ Normal	□ Sarcasm □ Humor □ Swearing □ Quote
2	RT @StLouisNewsHeds: St. Louis (MO) Business Journal: Rick Perry in St. Louis today #STL	○ Relevant --> ○ Not Relevant	○ For --> ○ Against --> ○ Mixed ○ Neutral ○ Can't Tell	○ Passionate ○ Excited ○ Normal	□ Sarcasm □ Humor □ Swearing □ Quote

Figure 7.2: Labeling interface

Table 7.2: Annotator agreement as percentage label overlap

	Relev.	Sent.	Subj.*	Pol.*	Int.	Sarc.	Humor	Swear.	Quote
Bachmann	0.871	0.493	0.802	0.924	0.460	0.854	0.751	0.975	0.865
Gingrich	0.572	0.271	0.379	0.569	0.212	0.587	0.533	0.598	0.526
Cain	0.816	0.436	0.700	0.807	0.452	0.762	0.677	0.845	0.758
Perry	0.792	0.436	0.652	0.784	0.420	0.782	0.654	0.819	0.733
Romney	0.826	0.446	0.642	0.817	0.404	0.861	0.828	0.872	0.850
Paul	0.794	0.374	0.651	0.616	0.478	0.791	0.737	0.828	0.733
Santorum	0.726	0.407	0.560	0.710	0.369	0.728	0.678	0.742	0.723
Average	0.793	0.425	0.660	0.776	0.416	0.784	0.701	0.839	0.760

7.2 Subset Analysis

Table 7.3 shows the relevance and sentiment statistics for each politician and their aggregates based on manually annotated data. One of the most striking features is high percentage of relevant documents – 94.8% on average. The accuracy of our retrieval method – using Twitter Search API using politician's names – supports the widespread use of this technique in the literature [19, 11, 84]. Looking at sentiment

Table 7.3: Subset relevance and sentiment statistics

				Sentiment (counts)					Sentiment (%)				
	total	relevant	rel %	for	aga	mix	neu	ctt	for	aga	mix	neu	ctt
Bachmann	1400	1295	92.5	115	728	31	317	105	8.9	56.2	2.4	24.5	8.1
Gingrich	600	575	95.8	69	218	20	160	108	12.0	37.9	3.5	27.8	18.8
Cain	1000	947	94.7	202	450	23	203	71	21.3	47.5	2.4	21.4	7.5
Perry	1400	1340	95.7	126	743	39	297	135	9.4	55.4	2.9	22.2	10.1
Romney	800	760	95.0	104	310	32	254	60	13.7	40.8	4.2	33.4	7.9
Paul	800	758	94.8	303	161	31	166	97	40.0	21.2	4.1	21.9	12.8
Santorum	400	389	97.3	76	150	19	128	16	19.5	38.6	4.9	32.9	4.1
Total/Avg	6400	6064	94.8	995	2760	195	1525	592	16.4	45.5	3.2	25.1	9.8

annotations, nearly 10% of the tweets were labeled as Can't Tell, with Gingrich having the highest %, and another 25% as Neutral, leaving 65% of the documents with subjective labels. We also observe the dominant sentiment class to be Against, with a notable exception of Ron Paul, whose For tweets outnumber Against nearly 2 to 1. Recall that in the Twitter/Reviews/Blogs study in Chapter 5 when we examined queries such as movies, cell phones, and restaurants, the sentiment distribution was quite different with 45% Neutral, 36% Positive and 7% Negative. The drastic difference between these statistics shows the uniqueness of political discourse.

Following [19], we summarize sentiment expressed in these documents by subtracting the number of Against tweets from For, resulting in a summary score of expressed sentiment. In general, the number of For and Against documents reflects the overall posting rate, but with Against at a greater rate, as shown in Figures 7.3 and 7.4. Again, Ron Paul is the only politician whose sentiment score is usually positive. Some other candidates do get a favorable sentiment score at the beginning

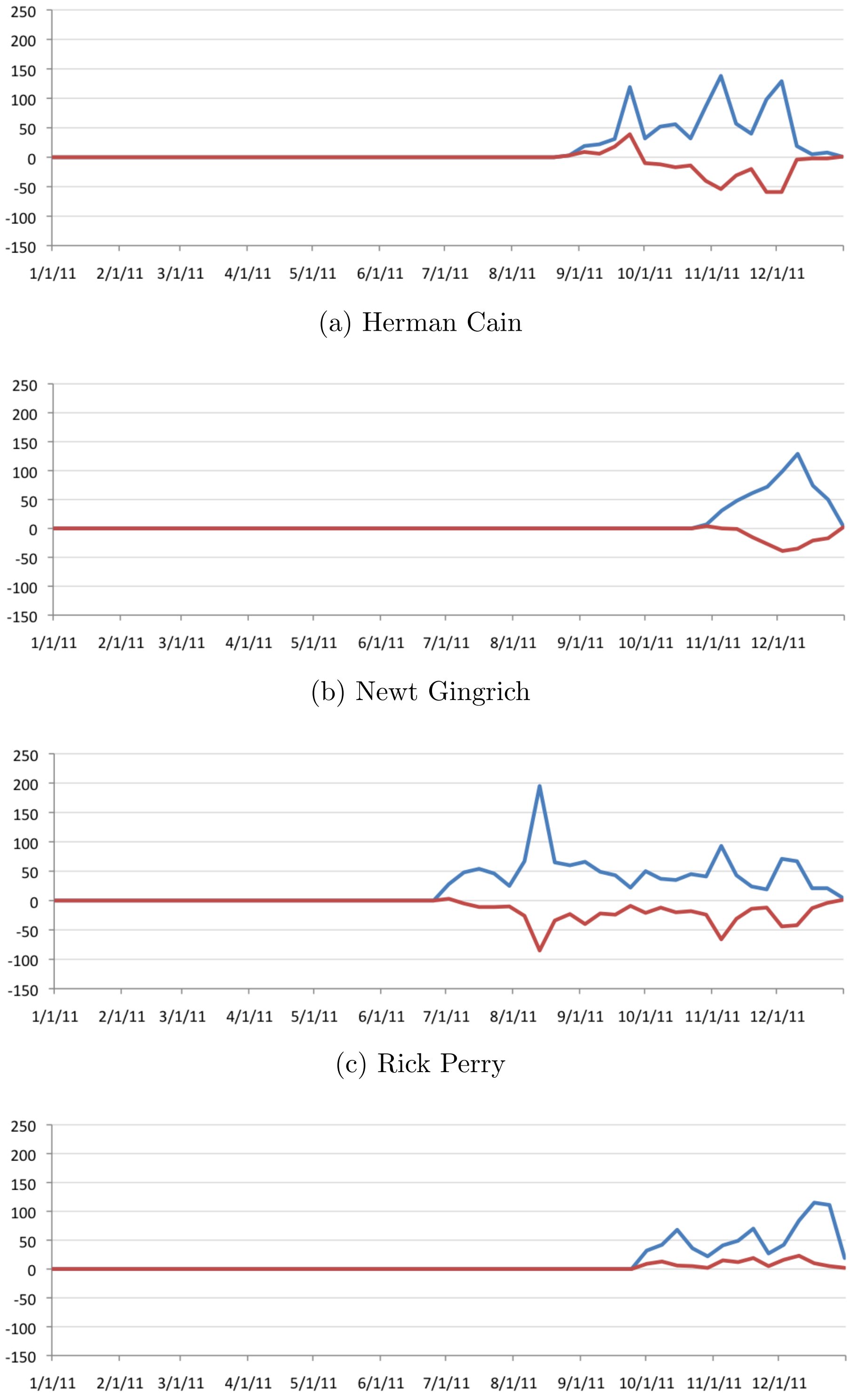
Table 7.4: Intensity associated with different sentiments (percent of total)

	For	Against	Mixed	Neutral	All
Passionate	8.3	6.6	3.5	0.9	6.6
Excited	19.5	15.2	18.8	2.2	15.4
Normal	72.2	78.2	77.6	97.0	78.0

of their campaigns, such as in september for Herman Cain. Otherwise, the sentiment score becomes negative and stays approximately the inverted scaled reflection of the volume line.

We also examine the intensity associated with each of the sentiments in Table 7.4. Tweets with Neutral sentiment show the least number of excited and passionate tweets, whereas those with For sentiment show the greatest. Note that a politically neutral tweet can still be excited, such as in ambiguous questioning: "*Did Michele Bachmann Jump the Shark by Suggesting HPV Vaccine Can Cause "Mental Retardation"? http://t.co/48zMcaN*". However, we find that on average 78% of the tweets are not particularly more intense than "normal". It would be interesting to compare levels of excitation with sentiment on other topics.

Finally, Table 7.5 shows the distribution of various stylistic features across tweets about each politician and in For and Against tweets separately. We find 21.6% of all tweets in the dataset to be humorous and 7.4% sarcastic. These are not evenly distributed between the politicians. For example, discussion about Bachmann and Cain are especially laden with sarcasm and humor. Strikingly, 40% of Against tweets



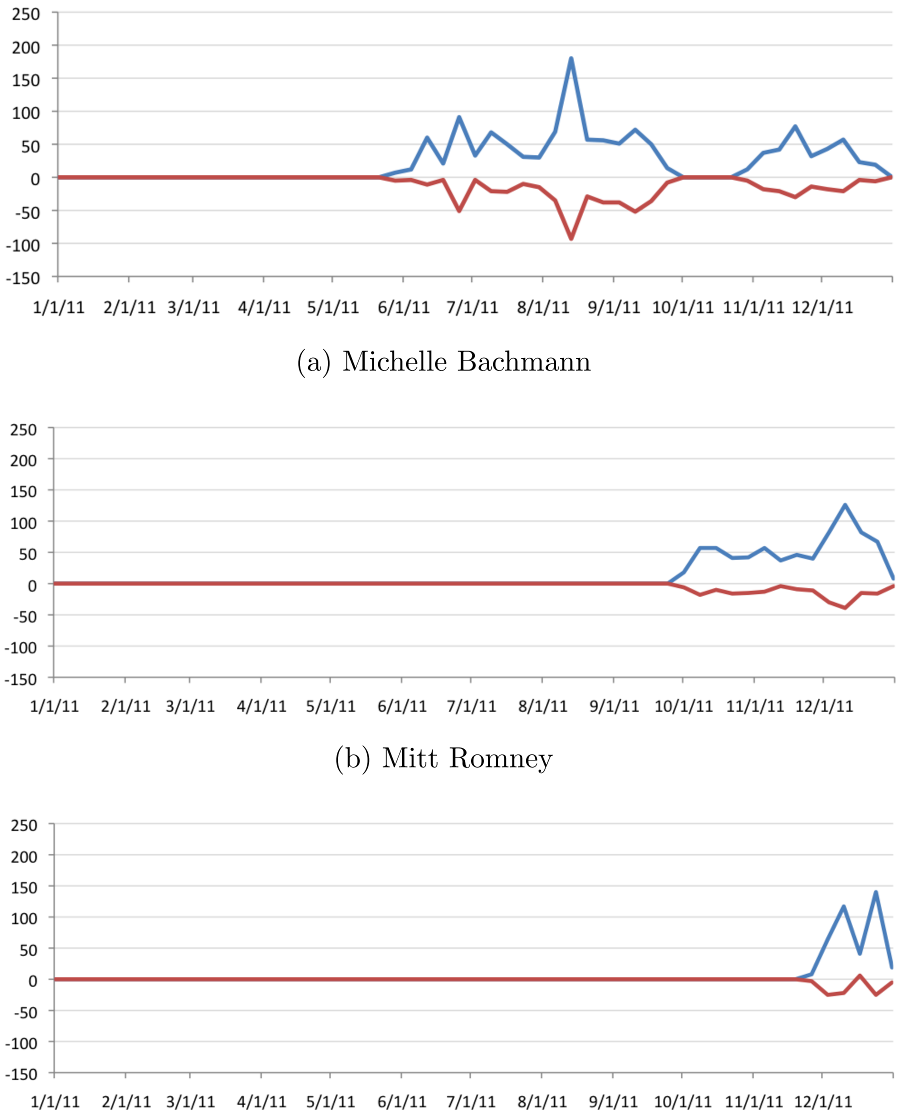
(a) Herman Cain

(b) Newt Gingrich

(c) Rick Perry

(d) Ron Paul

Figure 7.3: Weekly sentiment scores. Blue - total # tweets, red - sentiment score.



(a) Michelle Bachmann

(b) Mitt Romney

(c) Rick Santorum

Figure 7.4: Weekly sentiment scores. Blue - total # tweets, red - sentiment score.

Table 7.5: Stylistic features (percent of total)

	All Relevant				For				Against			
	sarc	humor	swear	quote	sarc	humor	swear	quote	sarc	humor	swear	quote
Bachmann	13.8	28.8	4.8	17.9	1.7	7.0	0.0	13.9	23.4	44.1	8.0	22.7
Gingrich	3.3	13.7	2.3	11.0	0.0	4.3	0.0	14.5	8.7	30.7	6.0	16.5
Cain	10.8	29.0	6.5	20.4	3.5	6.9	2.0	22.8	19.8	52.9	12.4	25.3
Perry	6.5	29.6	3.3	10.2	2.4	4.8	4.0	7.1	11.0	47.8	4.8	12.5
Romney	2.4	7.6	0.4	4.1	1.9	3.8	0.0	3.8	4.2	11.9	1.0	6.1
Paul	3.6	9.0	1.6	19.7	2.0	6.6	1.7	22.4	13.0	24.2	3.7	25.5
Santorum	3.9	15.9	1.8	4.9	1.3	2.6	0.0	5.3	8.7	37.3	4.7	5.3
All	7.4	21.6	3.3	13.6	2.1	5.7	1.4	15.8	14.7	40.3	6.5	17.2

are humorous, compared to only 5.7% of those For the politician. They are also more likely to contain swear words. Discussion about Mitt Romney shows much less of such rhetoric. We also note that a humorous tweet is 76.7% likely to also be sarcastic (but sarcastic tweets is only 26.2% likely to be humorous). This connection between sarcasm and humor would be an interesting future study. Our dataset contains 1311 humorous and 447 sarcastic documents – a dataset which could be used for such a study.

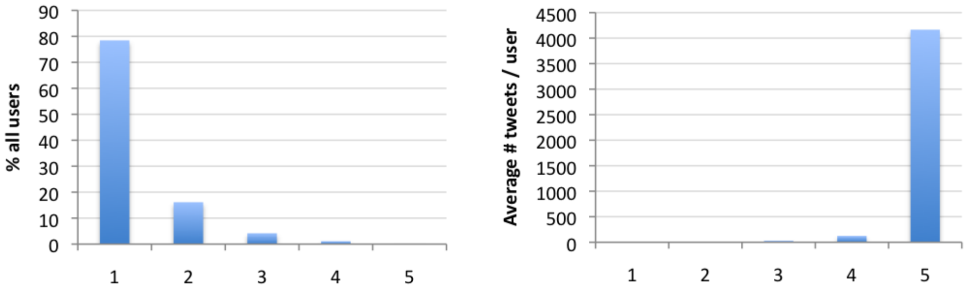
7.2.1 Silent Majority versus Vocal Minority

We further examine data by stratifying the users, as in Mustafaraj et al. [55]. We note that user posting behavior in our dataset follows power law – with few users posting thousands of messages and vast majority posting very few. Mustafaraj calls these extremes Silent Majority and Vocal Minority. Thus, we separate all users in our dataset into five quintiles according to their posting behavior (see Table 7.6). To do this, we separate the users into groups where each group is responsible for roughly

Table 7.6: Users grouped by posting behavior (in original dataset)

Group #	# of users	% of all users	tweets generated	% of all tweets	tweets/user
1	2,461,806	78.5	3,133,990	23.2	1.7
2	505,786	16.1	2,805,942	20.7	7.6
3	130,398	4.2	2,728,078	20.2	29.1
4	34,559	1.1	2,710,589	20.0	124.3
5	5,278	0.2	2,147,345	15.9	4164.5

a fifth of all content. The first group consists of 78.5% of all users in the dataset, in which users post an average of 1.7 tweets (that's over the span of a year). The most active group, however, consists of just 0.2% of all users, but it generated 15.9% of the tweets with an average of 4,164.5 tweets per user. The difference between the groups is illustrated in Figure 7.5 where the user membership is shown on the left and the average tweeting rate on the right. So, does the behavior of these groups differ?



(a) User membership

(b) Average tweeting rate

Figure 7.5: User group statistics

Table 7.7 shows sentiment and stylistic features of the tweets from each of

Table 7.7: Sentiment and stylistic features within stratified user groups

	Excluding Ron Paul					Ron Paul Only				
	1	2	3	4	5	1	2	3	4	5
For	10.2	10.8	10.5	12.7	19.3	33.3	29.0	33.5	44.1	48.5
Against	54.0	49.4	46.1	43.7	35.5	23.3	28.4	25.7	15.5	8.4
Mixed	2.2	3.0	3.7	2.6	3.3	5.3	7.1	2.4	4.3	0.6
Neutral	17.0	21.6	24.8	29.6	30.8	18.7	15.5	18.0	23.6	27.5
Can't Tell	9.2	9.0	10.2	7.8	7.8	11.3	13.5	12.6	9.3	13.8
Sarcasm	10.1	9.2	7.6	5.6	3.9	6.0	3.2	5.4	1.2	1.2
Humor	32.8	24.5	20.0	17.7	12.5	16.7	10.3	7.2	6.2	3.0
Swearing	5.7	3.9	2.8	3.0	0.9	3.3	1.3	0.6	1.9	0.6
Quotation	13.8	12.2	13.5	10.4	9.7	18.0	16.1	15.6	19.3	24.0
Hashtags	23.3	28.1	30.2	34.1	39.1	26.0	38.1	39.5	37.9	47.3
Links	43.0	51.8	57.7	60.4	63.5	42.0	43.2	55.1	72.0	81.4
Retweets	37.5	32.8	35.4	36.6	40.5	27.3	32.3	34.7	38.5	47.3
Only text	24.1	21.1	16.0	15.0	10.5	30.0	23.9	19.8	11.2	6.6

the user group. First note the bottom four characteristics extracted using regular expressions. The results are shown for Ron Paul and other politicians separately, because of the unusually positive overall sentiment of Mr. Paul's subset. We see many tendencies: the vocal group tends to be more for and less against the politician, and post more neutral tweets. It is also less sarcastic or humorous, but is more likely to use hashtags and links, and retweet. They are unlikely to post a tweet without any hashtags, links or retweet ("Only text" row). Ron Paul tweets show the same trends, except for the prominence of For sentiment.

Upon examining a selection of users from most and least vocal (around 70 users from each group), we note that whereas all users from the least vocal group were accounts owned by individuals (many of which had very few tweets), only 65% were individual accounts in the vocal group. These accounts have thousands of followers, and many have their own blogs or websites. Furthermore, 31% of the vocal group were campaigning for some political cause, and the last 4% were news sites.

The significance of these peculiarities is that the vast difference in posting frequency of these users skews the overall sentiment of the data. When polls measure favorability, each polled person is counted equally. This is not the case when each tweet is counted as a "vote". Thus, counting users instead of individual tweets may be a better approach when comparing sentiment expressed on Twitter to traditional polls.

In summary, our annotated dataset reveals a discussion 65% of which is opinionated speech, which is laden with humor and sarcasm. It showed the power-users to be more for the politician they are tweeting about, and to be less sarcastic, humorous, and use fewer swear words. The sample set also contains links (in 55.2% of sampled tweets), hashtags (31%), and retweets (36.4%). Compare these to a general subset we collected to estimate general Twitter use consisting of 5 million tweets, 13.0% of which had links, 16.5% had hashtags and 13.1% retweets. The opinionated speech is mostly biased against the politicians (except for the case of Ron Paul), and in which users with different posting behaviors exhibit different biases. These biases sometimes are also slightly positive at the beginning of the politician's candidacy. We explore the sentiment change in the sentiment tracking experiments later, which suggest that Twitter users may have a liberal (or anti-Republican) bias.

7.3 Classification

In this section, we develop a political sentiment classifier and evaluate it using our labeled dataset. Instead of using Lingpipe classifier as in previous experiments, we choose SVMlight[2], which gives us more flexibility for tuning of the class selection. Using Lingpipe tokenizer, we extract 1-, 2-, and 3-grams as a feature vector. Note that punctuation was not removed at this step and no stemming was performed on the words in order to capture twitter-specific features such as hashtags, mentions, as well as emoticons. Preliminary studies showed it to be beneficial to compute the models for the dataset as a whole instead of building one for each politician, and we take this approach. Classifiers were built to identify various features of the text – from relevance to individual sentiments. Performance of the classifiers (estimated using 10-fold cross-validation) is shown in Table 7.8 along with the majority baselines. For each task, for instance, Relevance, the table lists the two classes the classifier is meant to distinguish: Relevant and Not Relevant. The precision, recall, and F-measure are then shown for the two classes separately.

The performance of most classifiers shows to be just above those for the majority classes, with a notable exception of the classifier detecting Against tweets (notably, Lingpipe logistic regression classifier gives similar performance). Because we want to build a classifier which ultimately detects For, Against, and Neutral labels, we focus on improving the last three classifiers. Note that the recall values are very low for the minority classes (For and Neutral), thus we attempt to improve these.

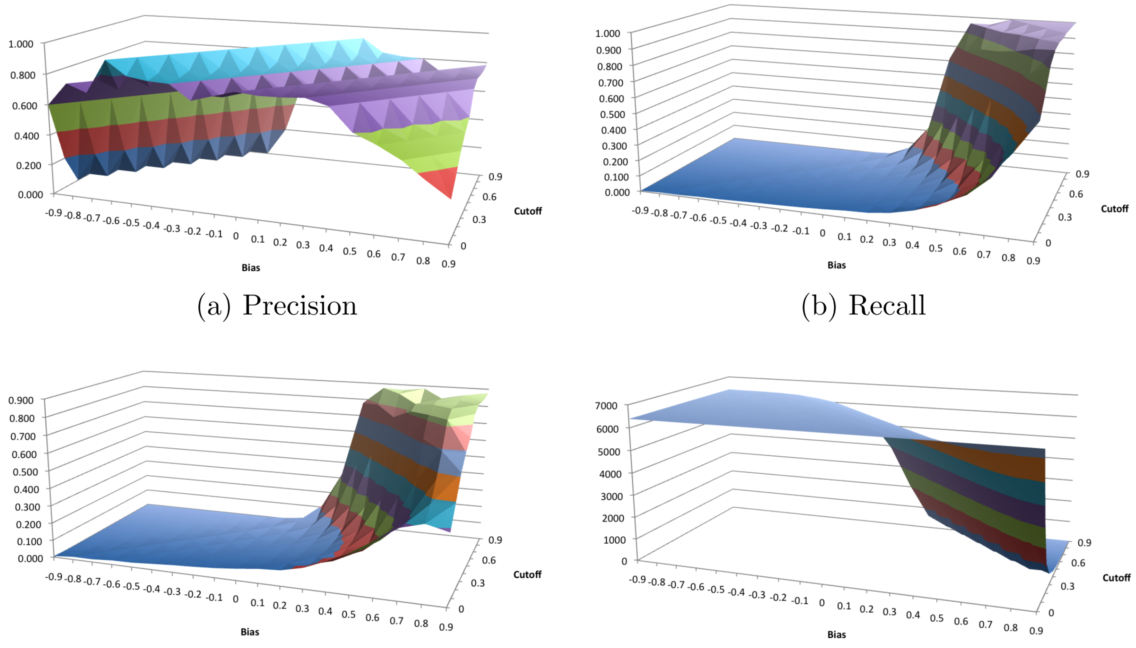
[2] http://svmlight.joachims.org/

Table 7.8: SVMlight classification performance for several tasks, with classes of documents the task is detecting listed as Class 1 and Class 2

				Class 1			Class 2		
	Class 1	Class 2	Accuracy	Prec	Rec	F	Prec	Rec	F
Relevance	Relevant	Not relevant	0.953	0.954	0.999	0.976	0.820	0.118	0.204
Majority			0.948	0.947	0.947	1.000	0.000	0.000	0.000
Subjectivity	Subjective	Objective	0.753	0.758	0.968	0.850	0.715	0.197	0.307
Majority			0.721	0.721	1.000	0.838	0.000	0.000	0.000
Polarity	For	Against	0.783	0.707	0.310	0.428	0.794	0.953	0.866
Majority			0.735	0.000	0.000	0.000	0.735	1.000	0.847
For	For	All others	0.849	0.798	0.037	0.071	0.849	0.998	0.918
Majority			0.845	0.000	0.000	0.000	0.845	1.000	0.916
Against	Against	All others	0.701	0.688	0.564	0.620	0.708	0.805	0.753
Majority			0.569	0.000	0.000	0.000	0.569	1.000	0.725
Neutral	Neutral	All others	0.775	0.715	0.095	0.167	0.777	0.988	0.870
Majority			0.762	0.000	0.000	0.000	0.762	1.000	0.865

To determine the class of a document SVMlight looks at the polarity of a score, which ranges roughly between -1 and 1. The magnitude of this score can be considered as the confidence of the classifier. Thus we introduce a notion of "cutoff", such that if the score is greater than the cutoff, the class decision is accepted. Furthermore, we may also want to change the value 0 as being the class cross-over point. We can "bias" the classifier by shifting this point closer toward -1 or 1. That is, if we change the cross-over point to -0.2, all documents in the range of [-0.2, 0.0] are now considered in the positive class instead of negative. We use a tuning set to determine the best values of the cutoff and bias by examining performance metrics at various values of these two parameters.

Precision, recall, F-measure, and number of classified documents for For classifier are plotted in Figure 7.6 (the distributions look very similar for Against and



(a) Precision

(b) Recall

(c) F-measure

(d) Number Classified

Figure 7.6: Performance metrics of For classifier with various cutoff and bias values

Table 7.9: SVMlight combined classifiers

Default SVMlight

	Accuracy	Avg Prec	Avg Rec	Avg F-measure
Overall performance	0.269	0.595	0.357	0.295
	Accuracy	Precision	Recall	F-measure
For	0.849	0.740	0.037	0.071
Against	0.578	0.680	0.556	0.612
Neutral	0.432	0.750	0.095	0.168
Other	0.475	0.211	0.738	0.328

Tuned SVMlight

	Accuracy	Avg Prec	Avg Rec	Avg F-measure
Overall performance	0.476	0.511	0.445	0.440
	Accuracy	Precision	Recall	F-measure
For	0.855	0.590	0.225	0.326
Against	0.594	0.681	0.553	0.610
Neutral	0.521	0.522	0.440	0.478
Other	0.632	0.251	0.561	0.347

Tuned SVMlight + Regression

	Accuracy	Avg Prec	Avg Rec	Avg F-measure
Overall performance	0.544	0.529	0.432	0.434
	Accuracy	Precision	Recall	F-measure
For	0.852	0.551	0.249	0.344
Against	0.592	0.565	0.836	0.674
Neutral	0.634	0.496	0.503	0.500
Other	0.826	0.503	0.139	0.218

Neutral classifiers). Notice as both bias and cutoff increase, fewer documents are being classified (d) (remember that documents under the cutoff are not classified), but recall (b) and f-measure (c) improve. That is, there is a tradeoff between how many documents we are willing to classify and performance. To obtain the final values for our classifiers we choose the best bias and cutoff values at which at least half of all documents are being classified. These are 0.2 cutoff and 0.8 bias for For and Neutral classifiers and 0.5 cutoff and 0.5 bias for Against classifier.

Using these three classifiers, we build a classifier to label data as belonging to one of these four classes: For, Against, Neutral, and Other. To do this, we combine outputs of our three classifiers, choosing the final class heuristically, using majority vote and confidence intervals. We first test it with the cutoff and bias set to 0 (default SVMlight classifier), and then with the tuned parameter values. Finally, instead of using heuristics for determining the final class label, we train a logistic regression classifier (using Weka) using the outputs of the three classifiers as features. The performance of these three classifiers is shown in Table 7.9.

We see a substantial improvement when the tuned cutoff and bias parameters are used, and further boost to overall accuracy when using regression to determine final class. Notice that in overall performance the Accuracy is computed for the final classifier instead of taking an average of individual class accuracies, showing the overall accuracy instead of by-class accuracy.

To further optimize our classifier we also tried anonymizing the data set by replacing the names of the target politicians with a bogus feature, but such approach did not yield a superior performance. It may be the case that politician's names have distinguishing qualities which aid the classifier in its task (Ron Paul, for example, is probably highly associated with the For class).

7.4 Tracking Sentiment

Using classifier developed in previous section we now track change in sentiment. We focus on the time spans around Republican debates taking place during the 2011.

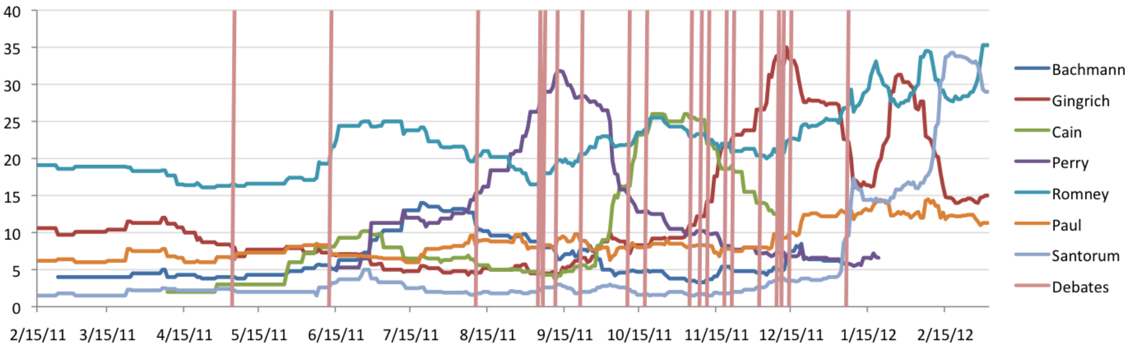


Figure 7.7: Polls for seven select politicians with debate days marked by vertical lines

A list of these debates was collected from 2012 Presidential Election News website[3]. We also collect the poll numbers for each of our politicians from Real Clear Politics[4], a website which collects information from national polls including Gallup, Rasmussen, Reuters, and others. Figure 7.7 shows the poll numbers for the seven politicians with vertical lines at the debates (19 in total).

For each debate, we collect a sample of 10,000 documents 5 days before and 5 days after the debate. The five-day window was chosen to accommodate the fact that the polls are not updated on a daily basis. We then apply SVMlight+Regression classifier to assign labels to the sampled documents. Our goal is to predict the change of sentiment that often happens around debates. We compare the change in predicted

[3] http://www.2012presidentialelectionnews.com/2012-debate-schedule/2011-2012-primary-debate-schedule/

[4] http://www.realclearpolitics.com/epolls/2012/president/us/republican_presidential_nomination-1452.html

class to that of the polls. Not all seven politicians participated at all of the polls, and our data did not cover some of the debates. The final experiment consisted of 104 predictions of sentiment change for a politician before and after debate. We take several approaches to estimating sentiment change:

- For: number of For documents after the debate minus before
- Against: number of Against documents before the debate minus after (reversed in order to show change in favorability)
- For-Against: number of For docs minus Against docs after the debate minus the same before
- For-Against Mod: same as For-Against, with For numbers boosted according to the average For to Against ratio (estimated using training set)
- * (U): same as above, but normalizing contribution of each tweet by the number of tweets the author has in the sample: $\frac{tweet\_polarity:\{-1,+1\}}{\#tweets\_by\_user}$
- Volume: number of all documents after the debate minus before

The performance of these approaches for each candidate is shown in Table 7.10. We also show the performance of a baseline based on the historical sentiment change in the polls: we predict sentiment change for a politician after a given debate according to the majority of sentiment changes in the previous debates for a that politician. For example, by the fourth debate in which Ron Paul participated, we have witnessed two debates after which the sentiment about him becomes more positive and one in which it becomes more negative, so we guess a positive change. Looking at prediction

Table 7.10: Predicting change in sentiment before and after debates: accuracy

	Bachmann	Gingrich	Cain	Perry	Romney	Paul	Santorum	Avg
For (T)	66.7	50.0	38.5	46.2	78.6	53.3	56.3	55.77
For (U)	60.0	55.6	30.8	53.8	78.6	60.0	68.8	58.65
Against (T)	53.3	44.4	69.2	61.5	28.6	40.0	50.0	49.04
Against (U)	66.7	38.9	61.5	61.5	42.9	40.0	50.0	50.96
For-Against (T)	60.0	50.0	69.2	61.5	28.6	46.7	50.0	51.92
For-Against (U)	60.0	55.6	69.2	53.8	28.6	40.0	56.3	51.92
For-Against Mod (T)	66.7	50.0	69.2	46.2	35.7	46.7	43.8	50.96
For-Against Mod (U)	60.0	50.0	69.2	53.8	35.7	40.0	50.0	50.96
Volume	53.3	66.7	46.2	46.2	57.1	73.3	50.0	56.73
Majority baseline	56.7	63.9	46.1	69.2	67.9	46.7	56.2	58.10

accuracy, we see different predictors performing differently for each politician. Ron Paul's sentiment change can be predicted quite well just by looking at the volume of conversation about him (which tends to be positive, unlike for the other candidates). Change in For and in Against documents showed different results. For example, the change in For documents predicts Romney sentiment change much better than the Against, but this is reversed for Herman Cain. Furthermore, normalizing the contribution of tweet sentiment by number of tweets posted by its user (U) increases the match for For and Against approaches. However, not any one of the approaches correlates well with the official poll results, and none are statistically better than the baseline. After computing Pearson correlation between these measures and the poll numbers, we also see very low numbers, with highest at 0.08.

We examine further the latest of the examined debates, one which took place on Jan 7, 2012. According to the polls, Gingrich did very poorly around the same time, seeing his numbers go from 27.4 (on Jan 2) to 16.6 (on Jan 11). The reverse is true for Santorum, whose numbers went from 4 (on Jan 2) to 15.8 (on Jan 11). First, we examine the top 10 most retweeted messages in our sample after the debate for Gingrich (warning: vulgarity):

A (39) *Newt Gingrich probably doesn't know how to use an ipod or eat his wife's *****.*

A (38) *Newt Gingrich is the poor man's Henry VIII, with his penchant for ditching wives & eating entire hams.*

A (28) *Gingrich's behavior is frankly so outside the norm for a Republican, only a Kenyan anti-colonial mindset can explain it.*

N (23) *New Hampshire: Romney 35, Paul 18, Huntsman 16, Gingrich 12, Santorum 11, Roemer 3, Perry 1: http://t.co/tlWlLG0x*

F (22) *RT our new video: "I want Newt Gingrich for President" http://t.co/Jn1lpVTP #withnewt*

A (21) *Todd Palin & Gingrich - a guy who thinks Alaska should secede from the Union endorses a guy who's seceded from two.*

A (21) *Romney deserves to be arrogant and isn't, while huntsman and gingrich don't deserve to be and are.*

A (20) *Newt Gingrich on leaving the race: "Not unless it gets cancer."*

A (19) *Ever since Peggy Noonan called Newt Gingrich "an angry little attack muffin" all I see is a screaming blueberry muffin when he talks.....*

A (19) *Debate grades, this time with all 6 candidates: Huntsman A, Santorum A-, Paul B, Perry B, Gingrich B-, Romney B-*

The majority of these popular tweets are anti-Gingrich jokes, with only one pro-Gingrich tweet. Also note that the popular jokes do not seem to be propagating because of an organized effort (such as in tweet supporting Gingrich – "RT our new video"), and they do not link to outside sources, but they are propagated just because the users thought they were worthy of sharing with others.

Lets look at Santorum's top 10 retweeted messages after the debate:

A (99) *Under Rick Santorum's health care plan, doctors will ask you to strip down to a sweater vest.*

A (66) *Rick Santorum's stance on homosexuality is so f***ing gay.*

A (28) *Rick Santorum was just introduced as the next president of the United States. Some in the crowd laughed. #fitn*

A (26) *Don't count Santorum out – He can still come from behind. #YeahIWentThere*

A (24) *Rick Santorum seems so homophobic that I'm surprised he even allows another man to vote for him.*

A (22) *New Study: Rick Santorum thinks about gay marriage more often than 79% of all gay men.*

A (22) *Defenders of biblical marriage blast Rick Santorum for his lack of slave wives and concubines.*

A (21) *Rick Santorum: "Life begins at conception." Mitt Romney: "Life begins at incorporation." Newt Gingrich: "Life begins when she gets cancer"*

A (21) *Dear Santorum, It's not just the "gay community who'd like to change laws" to grow equality. There's a group of folks called straight allies*

A (19) *Rick Santorum looks like the douchebag FBI agent whose inexperience gets everyone killed in an 80s action movie.*

For Santorum, the popular tweets also look quite bleak, with just about all of them jokes, and very few have Twitter-specific features which would make the tweet more searchable and retweetable (like hashtags, links, or pleas for users to retweet).

There are several directions for future experimentation available. First, we could check whether Twitter sentiment is predictive of or responsive to the national polls by "shifting" the times at which the sentiments are compared, for instance, by comparing earlier Twitter sentiment to later national polls. We may find a delayed response in Twitter to debates or other newsworthy events, but it would be even more interesting to find a sentiment which is first expressed in Twitter, and then in national polls. Second, we may examine polls which focus on a particular demographic, perhaps a younger population, or that which is more likely to express political opinion online. Similarly, there may be network characteristics which relate to sentiment, with “authorities” having more influence on Twitter sentiment.

The overall tendency towards the Against class in our dataset, as well as the fact that sentiment we find around the debates does not correspond well to that found in national polls suggests that political discourse on Twitter is not indicative of that of the nation as a whole. Because the politicians we examine in this project are Republicans (with possible exception of Ron Paul who has claimed to be Libertarian), it may be the case that an overall leaning of Twitter is more liberal. This may also be supported by Twitter’s young user base (mostly under 30)[5]. A future analysis of conversations about both republican and democrat politicians would shed more light on this issue.

7.5 Conclusion

In this project we analyze political discussion on Twitter about seven Republican candidates for US Presidential nomination over the year of 2011. We label a subset of this data for relevance, sentiment, sentiment intensity, and style, and examine it both as a whole, temporally, and by grouping users according to their posting behavior. We then build a multi-class classifier for identifying For, Against, and

[5] http://www.sysomos.com/insidetwitter/

Neutral sentiment and use it to track changes in sentiment around 19 debates.

7.5.1 Summary of Findings

We find that querying Twitter Search API using politician's names is an effective data gathering strategy, with 94.8% of the sampled documents relevant to their corresponding politicians. This is a popular technique already used (but sometimes not verified) in the literature [19, 11, 84]. The most striking feature of our subset was an overwhelming negative bias toward all politicians with an average ratio of 3.76 Against to 1 For tweets, except for Ron Paul who shows 0.53 to 1 ratio. The negative sentiment is sometimes matched by the positive at the beginning of the politician's campaign, but as a rule quickly returns to an overall negative sentiment. These negative documents are often humorous (40.3%) and/or sarcastic (14.7%), and sometimes contain swear words (6.5%).

By stratifying the users by the frequency of their postings, we find distinctly differing behaviors between the "silent majority" and "vocal minority" (terms coined by Mustafaraj et al. [55]). The vocal group tends to be more For and less Against the politician, it is less sarcastic and humorous, and is more likely to use hashtags, links, and retweet. Thus, if one counts users instead of tweets (as in traditional polls), the negative sentiment would be even more pronounced.

Using this dataset, we build and test a classifier to detect For, Neutral, and Against sentiments. We find that using out-of-the-box tools works nearly the same as the majority baseline, and only after some thorough tuning we improve overall accuracy from 0.269 to 0.544. We conclude that it is indeed a difficult task, and that researchers tracking political sentiment on Twitter should be wary of using untuned out-of-the-box tools without evaluation.

Using this classifier, we track sentiment expressed about each of the politicians change before and after 19 republican debates. We compare this sentiment to national polls, and find that overall the sentiment we find in the tweets does not well correspond to that in the polls. Examining the most popular tweets further, we find them mostly to be joking banter about the politicians, all negative – even for the politicians whose national poll numbers were improving. Such trends point to an overall anti-republican or liberal-leaning bias in the Twittersphere. Even the support we find for Ron Paul may be explained by the active young libertarian fan base for which he is famous[6]. Similar to our findings in the YouTube/Twitter study in the previous chapter, and within the limits of this study, we find that Twitter is a poor estimator of overall national political sentiment.

7.5.2 Future Work

This project is full of interesting future avenues of research. Does Twitter really have a liberal political bias? Could we detect such bias on an individual basis? Understanding the user base may bring us closer to understanding perhaps only a part of the national electorate. Tracking sentiments about politicians and issues from a range of political spectrum would show the preferences among and perhaps

[6] http://www.huffingtonpost.com/2012/01/12/ron-paul-young-voters_n_1202616.html

distinctions between twitter user base.

A further examination of the role of humor and sarcasm in political discourse could contribute to the design of sentiment labeling algorithms. Does sarcasm only reverse the apparent sentiment of the sentence, or is there a deeper semantic entanglement with sentiment? Also, which characteristics of the political topics become humorous, and could we predict how funny (and viral?) a statement may become? Our dataset contains 1311 humorous and 447 sarcastic tweets, which would be a good test bed for such analysis.

Finally, it may be useful to profile user behavior according to the frequency of their posts, the nature of their tweets, and the propagation of their network. We already show that users with different posting rate tend to differ in the polarity of their posts and several stylistic features. By doing this, we may not only find the users who are interesting or boring, but also potential spammers or bots.

CHAPTER 8
CONCLUSIONS

The five projects described in this thesis contribute answers to the following questions raised in our Introduction.

1. *Is it possible to enrich the definition of sentiment?*

First, we examine the very nature of sentiment, as it has been used in the field. The customary definition of sentiment as being negative or positive is a gross oversimplification of complex semantics of emotion. Thus, we introduce a model of affect developed in Sociology called Affect Control Theory (ACT). The empirically-derived multi-dimensional definition of sentiment it presents is not only more descriptive of affective meanings of individual words, but also provides higher-level formulations which combine individual meanings of words in a sentence to form a new summary meaning. We show that lexicons and analysis tools developed over the decades of ACT research are useful in the Sentiment Analysis task of polarity classification. Although the task considered here is a standard SA one, same techniques can be used to extract other ACT affective dimensions. Driven by this new definition of sentiment, further development of sentiment classification tools would benefit both text analysts and sociologists.

We further find that re-definition of sentiment may be necessary for effective sentiment analysis of complex topics like politics. In our comparison of YouTube and Twitter we show the subtle difference between positive/negative sentiment of writings and the author's agreement with the political stance. Emotions can be expressed in a variety of ways, with references to either side of an issue. For example, one may criticize policies of an opposing side when writing in support of another, thus expressing negative polarity while agreeing on topic in question. Thus, in the second project in which we examine political discourse we re-define sentiment as being either for, against, or neutral about the politician in question.

Because “sentiment” may mean different things in different kinds of discourse, one must not assume the customary definition applies. This means that tools designed to detect “sentiment” for one topic may be not applicable to a new topic. For example, we show in Chapter 6 that when applied to political speech general-purpose lexicon-driven polarity classifiers are no better than a majority-class baseline. Thus, as applications of sentiment analysis become more diverse, we show that new definitions of “sentiment” may be necessary, as well as the development of new tools.

2. *Which document representation approaches are the best for building data-driven sentiment classifiers?*

We next determine the best way to represent data in order to build data-driven sentiment classifiers. We test some of the latest popular feature definition, selection, and generalization techniques using three datasets of varying sizes and class memberships. We confirm some hypotheses, including that adjectives are important for polarity classification, and that stemming and using binary instead of term frequency feature vectors do not impact performance. We also show that the helpfulness of certain techniques depends on the nature of the dataset. For example, selecting top few features using Mutual Information hurts performance of the classifier on a smaller dataset, whereas it proves to be a good strategy for larger datasets. Cost analysis also reveals that some features used in the literature not only do not improve performance, but are very expensive to generate. In general, we observe that by combining low-order n-grams (n = 1, 2, 3) we achieve the best performance, and we use this approach in the following experiments. Finding this to be a strong approach, we conclude that a low-order n-gram classifier should be used as a baseline whenever a more complex algorithm is proposed.

3. *What are the differences and similarities between the expression of sentiment in different social media streams?*

Although a plethora of studies exists examining sentiment expressed in various social media sources, few rigorously compare sentiment across several data sources. In chapters 5 and 6 we build multi-stream topic-specific datasets, compare the sentiment expressed in these streams, and perform within- and cross-stream sentiment classification experiments. The Review/Twitter/Blog study focused on consumer-product-based topics such as movies and phones, whereas the YouTube/Twitter study examines political topics. In both cases, we find that although the proportion of positive to negative remained similar across streams, there were marked differences between them. Curiously, the class imbalance was quite different in the two studies, with texts about consumer-product-based topics mostly positive and political topics mostly negative. Such differences would be impossible to pick up using general samples, as have been previously examined [4].

Furthermore, in Review/Twitter/Blog study we show that classifiers built using reviews prove to be the most generalizable to other streams, followed by Twitter, with Twitter-based model performing as well as the native classifier 8 out of 10 for blogs and 5 out of 10 for reviews. We also show that combining training data from several streams further boosts performance, and combining data from different topics may even produce classifiers outperforming their native counterparts. However, the opposite is true for the political YouTube/Twitter collection, demonstrating the difficulty of classifying political speech. It may be the case that different behaviors (writing styles, attitudes, etc.) are captured in different streams, and at some point the models diverge too much. Yet such diversity may be necessary to classify other sources of political discourse, such as blogs and editorials, and we leave exploration of other sources of political writings to further research.

4. *What is political discourse in social media like, and is it indicative of national political sentiment?*

Finally, we examine political speech in two studies: one comparing YouTube comments to Twitter messages, and one tracking Republican politicians on Twitter. The most striking feature of our datasets was an overwhelming negative bias for all politicians (except for Ron Paul). The same negativity was expressed towards both republican and democrats (for example, for Barack Obama). The negative sentiment is sometimes matched by the positive at the beginning of the politician's campaign, but as a rule quickly returns to an overall negative sentiment. In Twitter, these negative documents are often humorous (40.3%) and/or sarcastic (14.7%), and sometimes contain swear words (6.5%).

Following [55], we examine the sentiment expressed by the majority of users who tweet very little and the minority who tweet a lot, and show that the vocal group tends to be more for and less against the politician, it is less sarcastic and humorous, and is more likely to use hashtags, links, and to retweet. Thus, if one counts users instead of tweets (as in traditional polls), the negative sentiment would be even more pronounced.

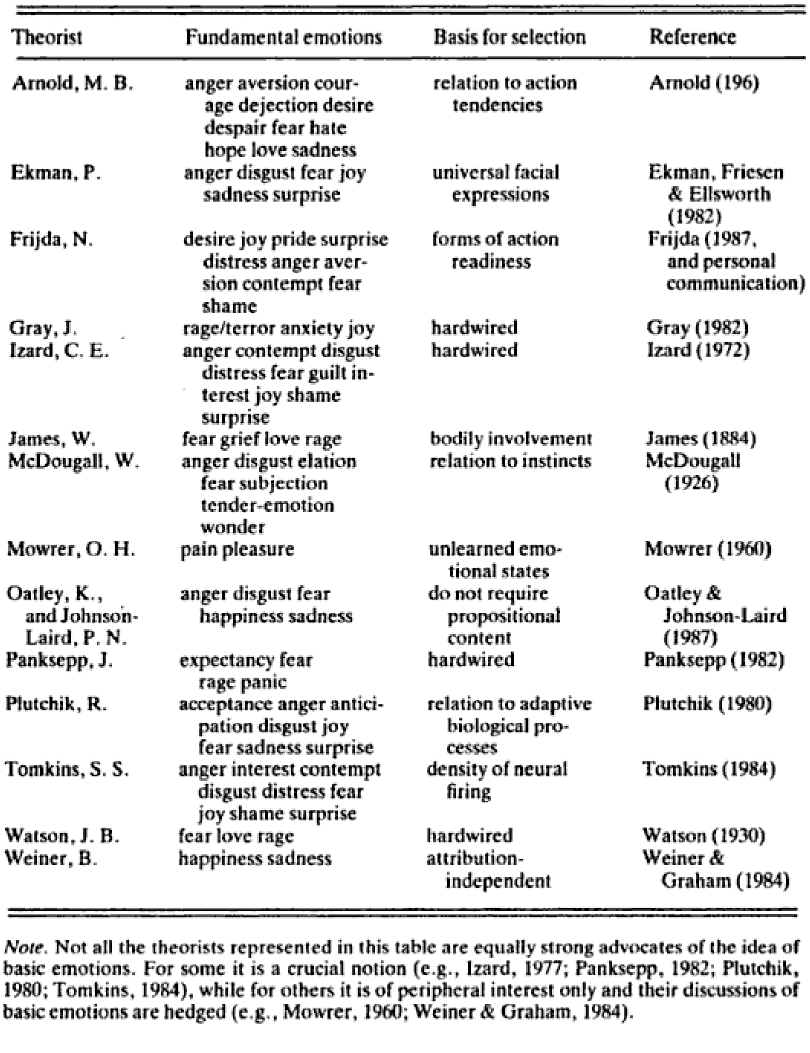
In both studies, we compare the sentiment in our datasets to that in national polls. In YouTube/Twitter study, we examine the volume of discussion and the amount of sentiment expressed about several politicians and find that these measures do not match well with the national Gallup polls. In the Twitter study, we examine the change in sentiment about a set of politicians before and after public debates and compare compare this sentiment to national polls. Again, we find that overall the sentiment we find in the tweets does not well correspond to that in the polls. Examining the most popular tweets further, we find them mostly to be joking banter about the politicians, all negative – even for the politicians whose national poll numbers were improving. Thus, we conclude that social media has a limited predictive power (if at all), as has been argued by [24, 53].

* * *

With the continuing evolution of social media and diversification of human expression online, text analysis tools need to adapt and develop to keep pace. Not only do the algorithms need to be re-examined, but the very definition of “sentiment” must be brought into question. The scale and ease of expression of emotion online is unprecedented, and it may be unwise to expect this expression to correspond to, say, standard opinion polls. As we show, political sentiment on YouTube and Twitter does not match national polls, yet with social media being such a powerful outlet for people's opinions, it must be examined in its own right. 2011 has been the year where social media has demonstrated its power to mobilize social movements as well as to share information about natural disasters, and as such, it is much more than an outlet for expressing opinions. We conclude that sentiment must be defined, and tools for its analysis designed, within the larger framework of human interaction.

APPENDIX A
DEFINING SENTIMENT

A selection of lists of “fundamental” or “basic” emotions



Theorist	Fundamental emotions	Basis for selection	Reference
Arnold, M. B.	anger aversion courage dejection desire despair fear hate hope love sadness	relation to action tendencies	Arnold (196)
Ekman, P.	anger disgust fear joy sadness surprise	universal facial expressions	Ekman, Friesen & Ellsworth (1982)
Frijda, N.	desire joy pride surprise distress anger aversion contempt fear shame	forms of action readiness	Frijda (1987, and personal communication)
Gray, J.	rage/terror anxiety joy	hardwired	Gray (1982)
Izard, C. E.	anger contempt disgust distress fear guilt interest joy shame surprise	hardwired	Izard (1972)
James, W.	fear grief love rage	bodily involvement	James (1884)
McDougall, W.	anger disgust elation fear subjection tender-emotion wonder	relation to instincts	McDougall (1926)
Mowrer, O. H.	pain pleasure	unlearned emotional states	Mowrer (1960)
Oatley, K., and Johnson-Laird, P. N.	anger disgust fear happiness sadness	do not require propositional content	Oatley & Johnson-Laird (1987)
Panksepp, J.	expectancy fear rage panic	hardwired	Panksepp (1982)
Plutchik, R.	acceptance anger anticipation disgust joy fear sadness surprise	relation to adaptive biological processes	Plutchik (1980)
Tomkins, S. S.	anger interest contempt disgust distress fear joy shame surprise	density of neural firing	Tomkins (1984)
Watson, J. B.	fear love rage	hardwired	Watson (1930)
Weiner, B.	happiness sadness	attribution-independent	Weiner & Graham (1984)

Note. Not all the theorists represented in this table are equally strong advocates of the idea of basic emotions. For some it is a crucial notion (e.g., Izard, 1977; Panksepp, 1982; Plutchik, 1980; Tomkins, 1984), while for others it is of peripheral interest only and their discussions of basic emotions are hedged (e.g., Mowrer, 1960; Weiner & Graham, 1984).

Source: **The Cognitive Structure of Emotions** by Ortony, Clore, and Collins. 1988.

APPENDIX B
AMAZON MECHANICAL TURK TWITTER HIT GUIDELINES

We are studying the emotional sentiment of a tweet's author to a particular topic. For this, we ask you to mark (1) whether a tweet talks about the topic, and if so, (2) the sentiment present in a tweet. The topics will range from movies, music albums, and computer games to mobile phones and restaurants. For example, the following tweets are marked for the topic of Movie: American Beauty:

- *"I loved watching American Beauty last night - so beautiful!"*
  - Is it about the topic? YES
  - What is the sentiment? Positive
- *"Yellowstone is a true American beauty!"*
  - Is it about the topic? NO
  - Then you don't need to fill the sentiment part.
- *"RT @fan2342 #americanbeauty #beauty #fitness"*
  - Is it about the topic? Can't Tell
  - Then you don't need to fill the sentiment part.
- *"Went to see American Beauty last night. The new theater is glam - soft seats, cup holders, you name it! But the movie was boring..."*
  - Is it about the topic? YES
  - What is the sentiment? Negative
- *"American Beauty: innovative plot, but the cast was a letdown... hmm..."*

  - Is it about the topic? YES
  - What is the sentiment? Mixed

- *"Come to the renovated theater downtown for a showing of American Beauty"*
  - Is it about the topic? YES
  - What is the sentiment? None

- *"@Angiexx RE: American Beauty – totally agree"*
  - Is it about the topic? YES
  - What is the sentiment? Can't Tell

If any of the tweets are not classified, the HIT will be discarded. One of the tweets is a control tweet (one with an obvious sentiment). If it is misclassified, the HIT will be discarded. There is also an 11th question, which must be answered to prove that you are not a bot.

Thank you for your contribution!

APPENDIX C
AMAZON MECHANICAL TURK BLOG POST HIT GUIDELINES

We are studying the emotional sentiment of a blog post's author to a particular topic. For this, we ask you to mark (1) whether the post talks about the topic, and if so, (2) which portion of the post talks about it, and finally (3) the sentiment present in that segment. The topics will range from movies, music albums, and computer games to mobile phones and restaurants. For example, the following blog posts are marked for the topic of Movie: American Beauty:

- *"Finally – a weekend! I cannot believe how much work my boss gave me. But I guess that's what we get for being a seasonal business. The pay will be nice as well! To celebrate, watched American Beauty with some friends last night. What a beautiful movie! Every scene is like a work of art!"*
  - Is it about the topic? YES
  - Relevant text: "watched American Beauty with some friends last night. What a beautiful movie! Every scene is like a work of art!"
  - What is the sentiment? Positive
- *"Yellowstone is a true American beauty! This is a shout out to all of my friends - we must get together this month and take a trip to one of the most dazzling American parks! Quickly, send me an email if you are interested. The current idea is to go for a few nights with camping."*
  - Is it about the topic? NO
  - Then you don't need to fill the relevant text or sentiment part.

- *”Click here for more on American Beauty — Hot Yoga — Welness Center — my favorite artists ”*
  - Is it about the topic? Can’t Tell
  - Then you don’t need to fill the relevant text or sentiment part.
- *”Went to see American Beauty last night. The new theater is glam - soft seats, cup holders, you name it! The ticket prices always surprise me though... I could get a decent meal for a movie theater ticket! Well, despite the nice seats I thought the movie was boring... I almost fell asleep a few times”*
  - Is it about the topic? YES
  - Relevant text: *”I thought the movie was boring... I almost fell asleep a few times”*
  - What is the sentiment? Negative
- *”American Beauty: innovative plot, but the cast was a letdown... hmm... Check out this review that I rather agree with: http://imdb.com/reviews/american-beauty/9498324”*
  - Is it about the topic? YES
  - Relevant text: *”American Beauty: innovative plot, but the cast was a letdown... hmm... ”*
  - What is the sentiment? Mixed
- *”Come to the renovated theater downtown for a showing of American Beauty. Student tickets* 5, *Adult7.”*
  - Is it about the topic? YES
  - Relevant text: *”a showing of American Beauty”*

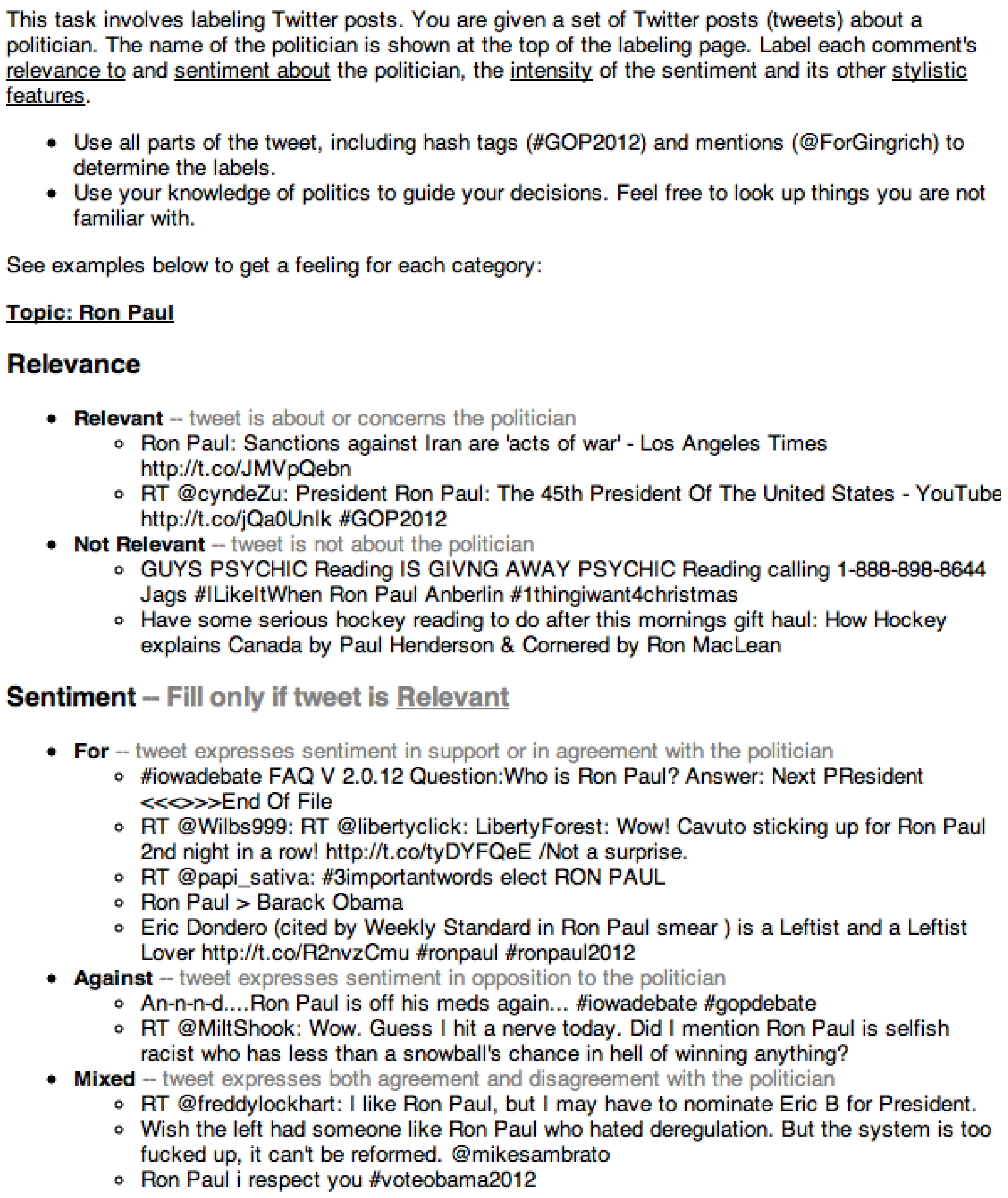
  - What is the sentiment? None

- *"Just read Joe's review of American Beauty on IMDB. I couldn't disagree more. Next time we meet, I'm debating this!"*
  - Is it about the topic? YES
  - Relevant text: *"Joe's review of American Beauty on IMDB"*
  - What is the sentiment? Can't Tell

If any of the blog posts are not classified, the HIT will be discarded. There is also a 6th question, which must be answered to prove that you are not a bot.

Thank you for your contribution!

APPENDIX D
POLITICAL SENTIMENT LABELING GUIDELINES



This task involves labeling Twitter posts. You are given a set of Twitter posts (tweets) about a politician. The name of the politician is shown at the top of the labeling page. Label each comment's relevance to and sentiment about the politician, the intensity of the sentiment and its other stylistic features.

- Use all parts of the tweet, including hash tags (#GOP2012) and mentions (@ForGingrich) to determine the labels.
- Use your knowledge of politics to guide your decisions. Feel free to look up things you are not familiar with.

See examples below to get a feeling for each category:

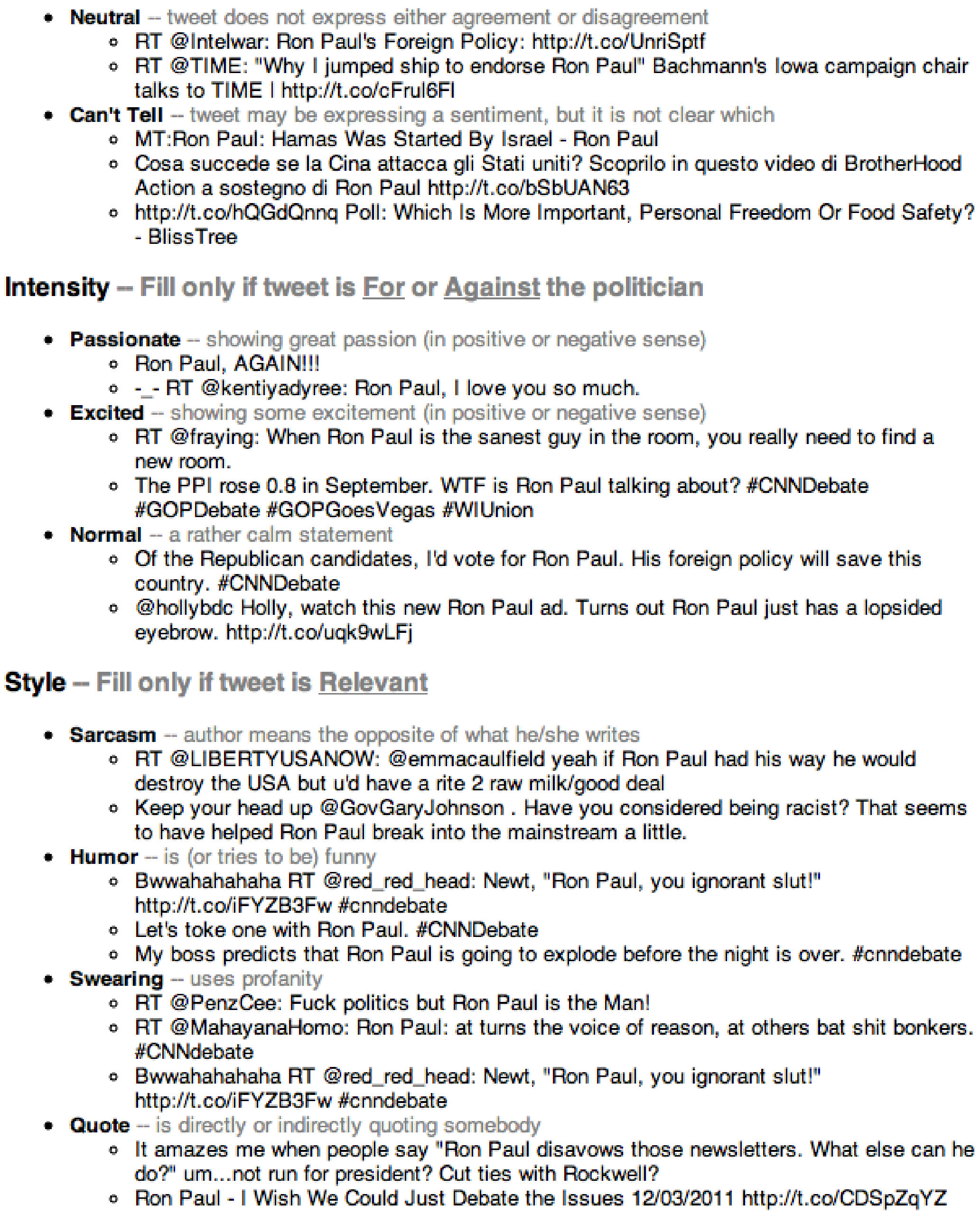
Topic: Ron Paul

Relevance

- **Relevant** -- tweet is about or concerns the politician
    - Ron Paul: Sanctions against Iran are 'acts of war' - Los Angeles Times http://t.co/JMVpQebn
    - RT @cyndeZu: President Ron Paul: The 45th President Of The United States - YouTube http://t.co/jQa0Unlk #GOP2012
- **Not Relevant** -- tweet is not about the politician
    - GUYS PSYCHIC Reading IS GIVNG AWAY PSYCHIC Reading calling 1-888-898-8644 Jags #ILikeItWhen Ron Paul Anberlin #1thingiwant4christmas
    - Have some serious hockey reading to do after this mornings gift haul: How Hockey explains Canada by Paul Henderson & Cornered by Ron MacLean

Sentiment -- Fill only if tweet is Relevant

- **For** -- tweet expresses sentiment in support or in agreement with the politician
    - #iowadebate FAQ V 2.0.12 Question:Who is Ron Paul? Answer: Next PResident <<<>>>End Of File
    - RT @Wilbs999: RT @libertyclick: LibertyForest: Wow! Cavuto sticking up for Ron Paul 2nd night in a row! http://t.co/tyDYFQeE /Not a surprise.
    - RT @papi_sativa: #3importantwords elect RON PAUL
    - Ron Paul > Barack Obama
    - Eric Dondero (cited by Weekly Standard in Ron Paul smear) is a Leftist and a Leftist Lover http://t.co/R2nvzCmu #ronpaul #ronpaul2012
- **Against** -- tweet expresses sentiment in opposition to the politician
    - An-n-n-d....Ron Paul is off his meds again... #iowadebate #gopdebate
    - RT @MiltShook: Wow. Guess I hit a nerve today. Did I mention Ron Paul is selfish racist who has less than a snowball's chance in hell of winning anything?
- **Mixed** -- tweet expresses both agreement and disagreement with the politician
    - RT @freddylockhart: I like Ron Paul, but I may have to nominate Eric B for President.
    - Wish the left had someone like Ron Paul who hated deregulation. But the system is too fucked up, it can't be reformed. @mikesambrato
    - Ron Paul i respect you #voteobama2012



- **Neutral** -- tweet does not express either agreement or disagreement
  - RT @Intelwar: Ron Paul's Foreign Policy: http://t.co/UnriSptf
  - RT @TIME: "Why I jumped ship to endorse Ron Paul" Bachmann's Iowa campaign chair talks to TIME | http://t.co/cFrul6Fl
- **Can't Tell** -- tweet may be expressing a sentiment, but it is not clear which
  - MT:Ron Paul: Hamas Was Started By Israel - Ron Paul
  - Cosa succede se la Cina attacca gli Stati uniti? Scoprilo in questo video di BrotherHood Action a sostegno di Ron Paul http://t.co/bSbUAN63
  - http://t.co/hQGdQnnq Poll: Which Is More Important, Personal Freedom Or Food Safety? - BlissTree

Intensity -- Fill only if tweet is For or Against the politician

- **Passionate** -- showing great passion (in positive or negative sense)
  - Ron Paul, AGAIN!!!
  - -_- RT @kentiyadyree: Ron Paul, I love you so much.
- **Excited** -- showing some excitement (in positive or negative sense)
  - RT @fraying: When Ron Paul is the sanest guy in the room, you really need to find a new room.
  - The PPI rose 0.8 in September. WTF is Ron Paul talking about? #CNNDebate #GOPDebate #GOPGoesVegas #WIUnion
- **Normal** -- a rather calm statement
  - Of the Republican candidates, I'd vote for Ron Paul. His foreign policy will save this country. #CNNDebate
  - @hollybdc Holly, watch this new Ron Paul ad. Turns out Ron Paul just has a lopsided eyebrow. http://t.co/uqk9wLFj

Style -- Fill only if tweet is Relevant

- **Sarcasm** -- author means the opposite of what he/she writes
  - RT @LIBERTYUSANOW: @emmacaulfield yeah if Ron Paul had his way he would destroy the USA but u'd have a rite 2 raw milk/good deal
  - Keep your head up @GovGaryJohnson . Have you considered being racist? That seems to have helped Ron Paul break into the mainstream a little.
- **Humor** -- is (or tries to be) funny
  - Bwwahahahaha RT @red_red_head: Newt, "Ron Paul, you ignorant slut!" http://t.co/iFYZB3Fw #cnndebate
  - Let's toke one with Ron Paul. #CNNDebate
  - My boss predicts that Ron Paul is going to explode before the night is over. #cnndebate
- **Swearing** -- uses profanity
  - RT @PenzCee: Fuck politics but Ron Paul is the Man!
  - RT @MahayanaHomo: Ron Paul: at turns the voice of reason, at others bat shit bonkers. #CNNdebate
  - Bwwahahahaha RT @red_red_head: Newt, "Ron Paul, you ignorant slut!" http://t.co/iFYZB3Fw #cnndebate
- **Quote** -- is directly or indirectly quoting somebody
  - It amazes me when people say "Ron Paul disavows those newsletters. What else can he do?" um...not run for president? Cut ties with Rockwell?
  - Ron Paul - I Wish We Could Just Debate the Issues 12/03/2011 http://t.co/CDSpZqYZ

REFERENCES

[1] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2010.

[2] Alexandra Balahur, Zornitsa Kozareva, and Andres Montoyo. Determining the polarity and source of opinions expressed in political debates. *Computational Linguistics and Intelligenct Text Processing*, 5449:468–480, 2009.

[3] Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato, and VS Subrahmanian. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. *International Conference on Weblogs and Social Media (ICWSM)*, 2007.

[4] Adam Bermingham and Alan Smeaton. Classifying sentiment in microblogs: Is brevity an advantage? *Conference on Information and Knowledge Management (CIKM)*, 2010.

[5] Albert Bifet and Eibe Frank. Sentiment knowledge discovery in twitter streaming data. *Lecture Notes in Computer Science*, 6332, 2010.

[6] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *Association of Computational Linguistics (ACL)*, pages 440–447, June 2007.

[7] J. Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *World Wide Web Conference (WWW)*, 2010.

[8] Francois-Regis Chaumartin. Upar7: A knowledge-based system for headline sentiment tagging. *SemEval-2007*, pages 422–425, 2007.

[9] P. Chesley, B. Vincent, L. Xu, and R. K. Srihari. Using verbs and adjectives to automatically classify blog sentiment. *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[10] Cynthia Chew and Gunther Eysenbach. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. *PLoS ONE*, 5(11), November 2011.

[11] Michael D. Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. *International Conference on Weblogs and Social Media (ICWSM)*, 2011.

[12] Hang Cui, Vibhu Mittal, and Mayur Datar. Comparative experiments on sentiment classification for online product reviews. *National Conference on Artificial Intelligence (AAAI)*, 21(2), 2006.

[13] S. Das and M. Chen. Yahoo! for amazon: Extracting market sentiment from stock message boards. *Asia Pacific Finance Association Annual Conference (APFA)*, 2001.

[14] H Daume. Frustatingly easy domain adaptation. *Association of Computational Linguistics (ACL)*, 2007.

[15] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *World Wide Web Conference (WWW)*, 2003.

[16] D. Davidov, Tsur O., and A. Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. *International Conference on Computational Linguistics (COLING)*, 2010.

[17] K Denecke. Using sentiwordnet for multilingual sentiment analysis. *Data Engineering Workshop, ICDEW*, pages 507 – 512, 2008.

[18] Ann Devitt and Khurshid Ahmad. Sentiment polarity identification in financial news: a cohesion-based approach. *Association of Computational Linguistics (ACL)*, pages 984–991, 2007.

[19] Nicholas A. Diakopoulos and David A. Shamma. Characterizing debate performance via aggregated twitter sentiment. *Conference on Human Factors in Computing Systems (CHI)*, 2010.

[20] Mark Dredze, Koby Crammer, and Fernando Pereira. Confidence-weighted linear classification. *International Conference on Machine Learning (ICML)*, 2008.

[21] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. *Language Resources and Evaluation (LREC)*, 2006.

[22] Albert Feller, Matthias Kuhnert, Timm O. Sprenger, and Isabell M. Welpe. Divided they tweet: The network structure of political microbloggers and discussion topics. *International Conference on Weblogs and Social Media (ICWSM)*, 2011.

[23] Michael Gamon, Sumit Basu, Dmitriy Belenko, Danyel Fisher, Matthew Hurst, and Arnd Christian Konig. Blews: Using blogs to provide context for news articles. *International Conference in Weblogs and Social Media (ICWSM)*, 2008.

[24] Daniel Gayo-Avello, Panagiotis Takis Metaxas, and Eni Mustafaraj. Limits of electoral predictions using twitter. *International Conference on Weblogs and Social Media (ICWSM)*, 2011.

[25] Jeffrey Ghannam. Social media in the arab world: Leading up to the uprisings of 2011. *A Report to the Center for International Media Assistance*, Febuary 2011.

[26] Sandra Gonzalez-Bailon, Rafael E. Banchs, and Andreas Kaltenbrunner. Emotional reactions and the pulse of public opinion: Measuring the impact of political events on the sentiment of online discussions. *http://arxiv.org*, 2010.

[27] E. Griffin. *A First Look at Communication Theory*. New York: The MacGraw-Hill Companies, 1997.

[28] David R. Heise. Affective dynamics in simple sentences. *Journal of Personality and Social Psychology*, 11:204–13, 1969.

[29] David R. Heise. *Understanding Events*. Cambridge University Press, 1979.

[30] Minquing Hu and Bing Liu. Mining and summarizing customer reviews. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005.

[31] Ziao Hu, J. Stephen Downie, and Andreas F. Ehmann. Lyric text mining in music mood classification. *International Society for Music Information Retrieval*, 2009.

[32] Valentin Jijkoun, Maarten de Rijke, and Wouter Weerkamp. Generating focused topic-specific sentiment lexicons. *Association for Computational Linguistics (ACL)*, pages 585–594, 2010.

[33] Litin Jindal and Bing Liu. Identifying comparative sentences in text documents. *Conference on Research and Development in Information Retrieval*, 2006.

[34] Nitin Jindal and Bing Liu. Review span detection. *WWW*, 2007.

[35] Mahesh Joshi and Carolyn Penstein-Rose. Generalizing dependency features for opinion mining. *Association for Computational Linguistics and Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, 2009.

[36] Nobuhiro Kaji and Masaru Kitsuregawa. Building lexicon for sentiment analysis from massive collection of html documents building lexicon for sentiment analysis from massive collection of html documents. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2007.

[37] Soo-Min Kim and Eduard Hovy. Crystal: Analyzing prediction opinions on the web. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

[38] Adam B. King. Affective dimensions of internet culture. *Social Science Computer Review*, 19:414–30, 2001.

[39] E. Kouloumpis, Wilson T., and J. Moore. Twitter sentiment analysis: The good the bad and the omg! *International Conference in Weblogs and Social Media (ICWSM)*, 2011.

[40] J.R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[41] Shoushan Li and Chengqing Zong. Multi-domain sentiment classification. *Human Language Technology (ACL HLT)*, 2008.

[42] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. *International Conference on Information and Knowledge Management (ACM CIKM)*, pages 375–384, 2009.

[43] Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann. Which side are you on? identifying perspectives at the document and sentence level. *Conference on Natural Language Learning (CoNLL)*, 2006.

[44] Yu-Ru Lin, James P. Bagrow, and David Lazer. More voices than ever? quantifying media bias in networks. *International Conference on Weblogs and Social Media (ICWSM)*, 2011.

[45] Bing Liu. *Web Data Mining*, chapter Opinion Mining. Springer, 2006.

[46] Huawen Liu, Lei Liu, and Huigie Zhang. Feature selection using mutual information: An experimental study. *PRICAI 2008: Trends in Artificial Intelligence. Lecture Notes in Computer Science*, 5351:235–246, 2008.

[47] Avishay Livne, Matthew Simmons, Eytan Adar, and Lada Adamic. The party is over here: Structure and content in the 2010 election. *International Conference on Weblogs and Social Media (ICWSM)*, 2011.

[48] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

[49] Tara McIntosh and James R. Curran. Weighted mutual exclusion bootstrapping for domain independent lexicon and template acquisition. *Proceedings of the Australasian Language Technology Workshop*, 6:97–105, 2008.

[50] Yelena Mejova. Tapping into sociological lexicons for sentiment polarity classification. *Young Scientists Conference, RuSSIR'10*, 2010.

[51] Yelena Mejova and Padmini Srinivasan. Exploring feature definition and selection for sentiment analysis. *International Conference in Weblogs and Social Media (ICWSM)*, 2011.

[52] Prem Melville, Woyciech Gryc, and Richard D. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. *Conference on Knowledge Discovery and Data Mining*, 2009.

[53] Panagiotis Metaxas, Eni Mustafaraj, and Daniel Gayo-Avello. How (not) to predict elections. *International Conference on Social Computing*, 2011.

[54] T. Mullen and N. Collier. Sentiment analysis using support vector machines with diverse information sources. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 412–418, 2004.

[55] Eni Mustafaraj, Samantha Finn, Carolyn Whitlock, and Panagiotis T. Metaxas. Vocal minority versus silent majority: Discovering the opinions of the long tail. *International Conference on Social Computing*, 2011.

[56] T. Nasukawa and J. Yi. Sentiment analysis: capturing favorability using natural language processing. *Proceedings of the Conference on Knowledge Capture*, 2003.

[57] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. *International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[58] Charles E. Osgood, W. H. May, and M. S. Miron. *Cross-cultural universals of meaning.* Urbana: University of Illinois Press, 1975.

[59] Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. *The Measurement of Meaning.* University of Illinois Press, 1957.

[60] Georgios Paltoglou and Mike Thelwall. A study of information retrieval weighting schemes for sentiment analysis. *Association for Computational Linguistics (ACL)*, pages 1386–1395, 2010.

[61] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Association for Computational Linguistics (ACL)*, 2004.

[62] Bo Pang and Lillian Lee. Thumbs up?: sentiment classification using machine learning techniques. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 10:79–86, 2002.

[63] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[64] Visqa Mani Kiran Peddinti and Prakriti Chintalapoodi. Domain adaptation in sentiment analysis of twitter. *Analyzing Microtext Workshop, AAAI*, 2011.

[65] John C. Platt. Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel Methods - Support Vector Learning*, 1998.

[66] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. *A comprehensive grammar of the English language*. Longman, 1985.

[67] Jacob Ratkiewicz, Michael D. Conover, Mark Meiss, Bruno Goncalves, Alessandro Flammini, and Filippo Menczer Menczer. Detecting and tracking political abuse in social media. *International Conference on Weblogs and Social Media (ICWSM)*, 2011.

[68] Ellen Riloff, Siddharth Patwardhan, and Janyce Wiebe. Feature subsumption for opinion analysis. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2006.

[69] Ellen Riloff and W. Phillips. An introduction to the sundance and autosslog systems. *Technical Report UUCS-04-015, School of Computing, University of Utah*, 2004.

[70] Dawn T. Robinson and Lynn Smith-Lovin. *Contemporary Social Psychological Theories*, chapter Affect Control Theory. Stanford Social Sciences, 2006.

[71] Allen Rubin and Earl R. Babbie. Research methods for social work. *Cengage Learning*, January 2010.

[72] Diego Saez-Trumper, Wagner Meira, and Virgilio Almeida. From total hits to unique visitors model for election's forecasting. *International Conference on Web Science*, 2011.

[73] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. *World Wide Web Conference (WWW)*, 2010.

[74] V.S. Sheng, F. Provost, and P.G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. *Conference on Knowledge Discovery and Data Mining (KDD)*, 2008.

[75] Lynn Smith-Lovin. Impressions from events. *Journal of Mathematical Sociology*, 13:71–101, 1987.

[76] Lynn Smith-Lovin and William Douglas. An affect control analysis of two religious subcultures. *Social Perspective in Emotions*, 1:217–48, 1992.

[77] R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2008.

[78] Swapna Somasundaran and Janyce Wiebe. Recognizing stances in ideological online debates. *NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 2010.

[79] Swapna Somasundaran, Theresa Wilson, Janyce Wiebe, and Veselin Stoyanov. Qa with attitude: Exploring opinion type analysis for improving question answering in on-line discussions and the news. *International Conference on Weblogs and Social Media (ICWSM)*, 2007.

[80] Carlo Strapparava and Alessandro Vlitutti. Wordnet-affect: and affective extension of wordnet. *International Conference on Language Resources and Evaluation (LREC)*, 2004.

[81] P. Subasic and A. Huettner. Affect analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy Systems*, (9):483–496, 2001.

[82] J. Suler. The online disinhibition effect. *CyberPsychology and Behavior*, 7:321–326, 2004.

[83] Songbo Tan, Zueqi Cheng, Yuefen Wang, and Hongbo Xu. Adapting naive bayes to domain adaptation for sentiment analysis. *Advances in Information Retrieval*, 5478:337–349, 2009.

[84] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Association for the Advancement of Artificial Intelligence Conference (AAAI)*, 2010.

[85] Mitch Wagner. Obama election ushering in first internet presidency. http://www.informationweek.com/news/government/212000815, 2008.

[86] X Wan. Co-training for cross-lingual sentiment classification. *Association for Computational Linguistics and Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, 2009.

[87] C. Whitelaw, N. Garg, and S Argamon. Using appraisal groups for sentiment analysis. *Conference on Information and Knowledge Management (CIKM)*, 2005.

[88] J. M. Wiebe. Tracking point of view in narrative. *Computational Linguistics*, 20:233–287, 1994.

[89] J. M. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin. Learning subjective language. *Computational Linguistics*, 30, 2004.