

How Reliable Is Sentiment Analysis? A Multi-domain Empirical Investigation

Tao Ding^(✉) and Shimei Pan^(✉)

University of Maryland, Baltimore County, Baltimore, MD 21250, USA
{taoding01,shimei}@umbc.edu

Abstract. Sentiment analysis (also known as opinion mining) is frequently used in monitoring public opinions on the internet. For example, it can help marketers evaluate the success of an ad campaign. It can also be used to assess public opinions during a political campaign. As a result, many businesses and organizations are exploring the potential value of employing sentiment analysis as a part of their business and social intelligence strategies. However, the technology isn't fully mature yet. As a result, if not used carefully, the results from sentiment analysis can be misleading. In this paper, we present an empirical investigation of the effectiveness of using current sentiment analysis tools to assess people's opinions in five different domains. The results were very uneven, from decent (e.g., hotel reviews) to poor (e.g., comments on public policies). We also proposed several *effectiveness indicators* that can be used to signal the appropriateness of using these tools in specific domains.

Keywords: Content analysis · Sentiment analysis · Performance measure

1 Introduction

With the rise of the World Wide Web, people are expressing their opinions and thoughts online using review sites, blogs, forums, and social networking sites. They collectively represent a rich source of information on different topics. Being able to capture the emotional responses of the public can help gain insight and make informed decisions. For example, it can help determine if a marketing initiative is driving the planned responses, or determine whether consumers prefer a new product just launched or not, or people's reaction to a political debate [5, 20]. To meet this need, many open source and commercial sentiment analysis (SA) tools have been developed. With these tools, more and more businesses, organizations, and individuals can harness the power of sentiment analysis by applying these tools directly to their data. Moreover, the easy availability of massive amount of opinion-rich online data also fuels the wide adoption of SA tools. For example, open-source web crawlers can be used to collect the review data easily. Many social media sites also release their application programming interfaces (APIs), which makes data collection from social media convenient.

Nowadays, SA has been widely used to gauge public opinions towards products [6], services [17], social events [23], political events [5], political candidates, and public policies [4, 20].

However, due to the complexity in automated text analysis, today’s sentiment analysis tools are far from perfect. For example, many of them are good at detecting useful mood signals (e.g., positive or negative sentiment) but inadequate in tracking and inferencing the relationships between different moods and different targets. As a result, if not used carefully, the results from sentiment analysis can be meaningless or even misleading. Since the typical users of SA are not researchers but business owners or individuals, they may not have the necessary knowledge to determine whether a SA tool is appropriate for their application domains or not.

In this paper, we present an empirical analysis of the effectiveness of using existing sentiment analysis tools for different applications. We have collected data from five different domains: movie reviews, hotel reviews, public comments on net neutrality, Tweets about political candidates, and public comments on Harvard University’s admission policy. Based on these data, we study the relations between the results of sentiment analysis and the corresponding common perception of the public opinion. To help determine whether a SA tool is appropriate for one’s data, we also proposed several *effectiveness indicators* that can be computed efficiently from given datasets.

The main contributions of our work include:

1. This is the first formal study known to us that analyzes the appropriateness of using sentiment analysis on diverse data sets. Our results can shed lights on the limitations of existing tools. Our results can also help raise the awareness of the potential pitfalls associated with the misuse of this technology.
2. We also propose a diverse set of *effectiveness indicators* that can be computed efficiently from given datasets to help people determine the appropriateness of using a sentiment analysis tool on given datasets.

In the following, we first review the current sentiment analysis methods and their applications, followed by a description of our datasets and the analyses we performed to assess the effectiveness of applying sentiment analysis on these datasets. Then we explain our effort in developing a few effectiveness indicators to help users determine whether a SA tool is appropriate for a given dataset. Finally, we conclude the paper by summarizing the main findings and pointing out a few future directions.

2 Related Works

Sentiment Analysis, also called opinion mining, in a broad sense is defined as the computational study of opinions, sentiments and emotions expressed in text [12]. According to [9], the task of sentiment analysis is to automatically extract a quintuple from text:

$$(e_i, a_{ij}, s_{ijkl}, h_k, t_l),$$

where e_i is a target object, a_{ij} is an aspect or attribute of e_i , s_{ijkl} is the sentiment value of aspect a_{ij} of entity e_i , h_k is the opinion holder, and t_l is the time when an opinion is expressed by a opinion holder. Once the sentiment quintuples are extracted from text, they can be aggregated and analyzed qualitatively or quantitatively to derive insights. Extracting the quintuples from unstructured text however is very challenging due to the complexity in natural language processing (NLP). For example, a positive or negative sentiment word may have opposite orientations in different application domains; Sarcasm is hard to detect; Coreference resolution, negation handling, and word sense disambiguation, a few well known but unsolved problems in NLP need for correct inference. Since many of the existing sentiment analysis tools did not solve these problems appropriately, they may work well in simple domains but not effective for more complex applications.

In terms of the methods used in typical sentiment analysis systems, they can be divided into lexicon-based and machine learning-based [10]. Since a purely lexicon-based approach is less common these days, here we focus on machine learning-based methods. Frequently, a machine learning-based system also incorporates lexical features from sentiment lexicons in its analysis.

Machine learning-based sentiment analysis can be further divided into supervised and unsupervised learning methods. The supervised methods make use of a large number of annotated training examples to build a sentiment classification model. Typical classification methods include Naive Bayes, Maximum Entropy classifiers and Support Vector Machines [13]. In general, for supervised sentiment analysis, if the target domain is similar to the source domain from which the training examples are collected, the prediction accuracy will be similar to the specified performance. In contrast, if the target domain is very different from the source domain, the sentiment analysis performance can deteriorate significantly. Among existing supervised sentiment analysis tools, some provide pre-trained models such as the Mashape Text-Processing API¹, others require users to provide labeled data and then train their own prediction models, such as Google Prediction API², NLTK text classification API³.

Since annotating a large number of examples with sentiment labels can be very time consuming, there are also many unsupervised sentiment analysis systems that do not require annotated training data. They often rely on opinion bearing words to perform sentiment analysis [1, 22]. Turney [19] proposed a method that classifies reviews by using two arbitrary seed words – poor and excellent, to calculate the semantic orientations of other words and phrases. Read [16] proposed a weakly-supervised technique, using a large collection of unlabeled text to determine sentiment. They used PMI [19], semantic spaces, and distributional similarity to measure similarity between words and polarity

¹ <http://text-processing.com/docs/sentiment.html>.

² <https://cloud.google.com/prediction/docs>.

³ <http://www.nltk.org/api/nltk.classify.html>.

prototypes. The results were less dependent on the domain, topic and time-period represented by the testing data. In addition, Hu [7] investigated whether models of emotion signals can potentially help sentiment analysis.

So far, hundreds of commercial state-of-the-art tools are available for automatic sentiment analysis, such as Semantria⁴, SentimentAnalyzer⁵, SentiStrength⁶, MLAnalyzer⁷, TextProcessing⁸. These tools can be applied directly to unlabeled documents without the need for domain-specific model training. In our experiment, we used Semantria as an unsupervised sentiment analysis tool to evaluate its effectiveness on different domains. Since most supervised sentiment analysis tools did not provide the original training data, we choose TextProcessing as a supervised sentiment analysis tool in our experiment since the original training data is available, which are movie reviews created by Pang [11]. As a result, the similarity between trained domain and target domains can be computed.

Fewer open-source tools dedicated to sentiment analysis are available today. To compare the results among different supervised methods, we train our Naive Bayes classifier using the NLTK API. The training data are the same as those in TextProcessing. To compare unsupervised tools, we employed SANN⁹ [14]. Table 1 summarizes the tools used in this investigation.

Table 1. Categorization of Selected tools.

Method	Tool
Supervised	NLTK API (Naive Bayes)
	TextProcessing
Unsupervised	SANN
	Semantria

3 Data Collection

To evaluate the impact of domain differences on sentiment analysis, we included five datasets:

1. **Hotel Reviews (Hotel)**: the dataset was originally used in [21]. We chose this dataset because reviews such as product reviews, hotel reviews and restaurant reviews are the most typical domains for sentiment analysis. In our study, we included 18726 reviews for 152 hotels, each includes the textual content, the author, and the overall rating that ranges from 1 star to 5 stars.

⁴ <https://semantria.com/>.

⁵ <http://sentimentanalyzer.appspot.com/>.

⁶ <http://sentistrength.wlv.ac.uk/>.

⁷ <https://www.publicapis.com/mlanalyzer>.

⁸ <http://text-processing.com/demo/sentiment/>.

⁹ <https://github.com/nik0spapp/unsupervisedsentiment>.

2. **Net Neutrality(NN):** The US Federal Communications Commission (FCC) [3] has published the public comments they received on the Open Internet/Network Neutrality bill. This bill considers the protection and Promotion of the principle of Open Internet to ensure that government and internet service providers should treat all data on the internet the same, not discriminating or charging differentially by user, content, site, platform, application, type of attached equipment, or mode of communication (FCC 14–28¹⁰). In our experiments, we included 26282 comments from this dataset. With this dataset, we want to evaluate the effectiveness of using sentiment analysis to assess public opinions towards a public policy.
3. **Tweet:** We collected a set of tweets related to the 2016 presidential campaign of Hillary Clinton. We used the search keywords “Hillary Clinton president” as the query to collect related tweets using the Twitter API. After filtering out redundant tweets, our dataset includes 7237 tweets. With this dataset, we want to investigate the effectiveness of using sentiment analysis to assess public opinions towards a political candidate based on social media posts since nowadays, social media-based opinion analysis becomes increasingly more popular.
4. **Harvard university Admission Policy (HAP):** In 2015, *Wall Street Journal* published an article on a lawsuit filed by a group of Asian-American organizations alleging that Asian-Americans face discriminatory standards for admission to Harvard University [2]. The complaint claimed that Harvard has set quotas to keep the number of Asian-American students admitted to the university much lower than their applications should warrant. We collected 924 public comments on this article. With this dataset, we want to study the effectiveness of using sentiment analysis to assess the public reaction toward a social event.
5. **Movie Review:** To investigate the impact of domain difference on the effectiveness on a supervised sentiment analyzer, we also include a dataset of movie reviews. The data source was the Internet Movie Database (IMDb). These reviews were originally used by Pang et al. (2002). They selected reviews where the author rating was expressed with stars. Ratings were automatically extracted and converted into one of three categories: positive, negative, or neutral. They only kept 1000 positive reviews and negative reviews for sentiment classification. Some existing sentiment analysis tool, such as TextProcessing, used these polarity data to train sentiment classifier. We compare the other four domains with the movie domain in our experiments to study the performance of supervised sentiment analysis tools.

Table 2 shows some statistics of these dataset.

¹⁰ <https://www.fcc.gov/rulemaking/most-active-proceedings>.

Table 2. Dataset.

	# of doc	# of sentence	size of corpus
Hotel	18726	171231	867795
NN	26282	88039	4672959
Tweet	7237	10160	867795
HAP	924	3105	25198
Movie	2000	64720	636524

3.1 Annotation Task

To evaluate the effectiveness of each sentiment analysis tool on different domains, we obtained two types of ground truth (1) the *emotion or feeling* expressed in a sentence or a message (called *emotion ground truth*) (2) the *opinion* expressed in a message (called *opinion ground truth*). Here we differentiate *emotion/feeling* from *opinion*. Emotion or feeling, is an immediate, instinctive and direct response to experience while opinion is more complicated. It is a combination of our autonomic emotional responses, behavior as well as cultural or societal meaning towards a subject. We would like to investigate whether the emotions or feelings expressed in a text is easier to detect than opinions since emotion is more direct while opinion is often indirect and appeals to preconceived notions and cultural norms.

To obtain the *emotion* and *opinion* ground truth, we used Amazon’s Mechanical Turk (AMT). Amazon Mechanical Turk is a crowdsourcing Internet marketplace that enables individuals and businesses (known as Requesters) to coordinate the use of a large number of workers (a.k.a Turkers) to perform tasks. In this case, we asked each Turker to read a post and decide the emotion and opinion expressed in the text. The emotion annotation is at both sentence and message level while the opinion annotation is only at the message level. For emotion annotation, each sentence or message is annotated with four labels: positive, negative, neutral and don’t know. The opinion label is specific for each application. To ensure the quality of the ground truth data, each post is annotated by three different annotators. All the annotators also have to be pre-qualified based on the following criteria: they must have submitted over 5000 tasks on AMT with an acceptance rate of over 95%.

Specifically,

For opinion annotation for hotel reviews, we ask each participant to decide whether 1. the author likes the hotel; 2. the author dislikes the hotel; 3. the author is neutral; 4. the author’s opinion is unclear.

One example from the hotel domain is the following:

Great Hotel Fantastic Hotel. Get the goldfish to keep you company. We still miss ours, Phil! Jeff at the concierge was a great help. Loved the crazy room—somehow the stripes work. Will definitely return. Breakfast at the restaurant was outstanding.

For emotion annotation, we first ask each participant to choose an emotion label for each sentence. After that, the participant also need to provide an overall emotion/feeling label for the entire post. Figure 1 shows the emotion annotation UI used in the AMT study. Overall 500 hotel reviews were selected randomly to be annotated on AMT.

For opinion annotation on the net neutrality public comment dataset, we asked each Turker whether 1. the author supports net neutrality; 2. the author is against net neutrality; 3. the author is neutral; 4. the author’s opinion is unclear. Then, the Turker was also asked to annotate emotions at both the sentence and the message level. Overall, 500 comments from the net neutrality dataset are selected randomly to annotate. Here is an example from the net neutrality dataset:

The Internet was created with public funds for the use of the public and the government. No for-profit organization should have the right to control access from the people who need and use it.

To annotate the opinions expressed in Twitter posts, we asked each Turker to rate whether 1. the author supports Hillary Clinton 2. the author does not support Hillary Clinton, 3. the author is neutral or 4. the opinion of the author is unclear. Then, the Turker is also asked to complete the emotion annotation task. We randomly selected 1000 tweets to annotate. Here is an example of such a tweet:

I WILL NOT vote for Hillary Clinton for President WE DO NOT want Bill BACK in the White House y’all know what I mean.

The HAP comments are more complex. Many contain deeply embedded conversation threads (e.g., comments on comments). In this case, sufficient context is particularly important for Turkers to understand the opinion expressed by different people. For example, one comment: *@David Smith: I totally agree with you, the university should pay attention to that.* is a reply to a previous comment expressed by David Smith. The opinion expressed in this comment is ambiguous if we don’t know the opinion of David Smith. To provide Turkers enough context to determine opinions expressed in a message, instead of providing a comment without context, we asked the Turkers to annotate an entire conversation thread. The following is a conversation thread from HAP:

Glenn Wilder : And of course the Dept Chair of African American Studies simply cannot be delivering lectures to a room full of Hispanics Asians and Caucasians. The class may actually have some value...but it would be lost on such a group. This alone justifies the need to balance out the student body.

Patrick O’Neil : @ Glenn Wilder This seems prejudicial! Why isn’t there a Chair of Hispanic American studies and Asian American studies?

Preston Moore : @ Glenn Wilder Don’t forget the Chair of the Women’s Studies dept or Chair of East Asia Languages.

2. What is feeling or emotion of customer 1 expresses in the sentence?

Great Hotel Fantastic Hotel.

positive 😊 neutral 😐 negative 😞 do not know

Get the goldfish to keep you company.

positive 😊 neutral 😐 negative 😞 do not know

We still miss ours, Phil!

positive 😊 neutral 😐 negative 😞 do not know

Jeff at the concierge was a great help.

positive 😊 neutral 😐 negative 😞 do not know

Loved the crazy room--somehow the stripes work.

positive 😊 neutral 😐 negative 😞 do not know

Will definitely return.

positive 😊 neutral 😐 negative 😞 do not know

Breakfast at the restaurant was outstanding.

positive 😊 neutral 😐 negative 😞 do not know

3. What is overall feeling or emotion of customer 1 in the comment?

positive 😊 neutral 😐 negative 😞 do not know

Fig. 1. An example of the AMT emotion annotation UI.

After reading each conversation thread, we ask each Turker to annotate the opinion and emotions expressed by each person involved in the conversation. For the above example, we ask each Turker to annotate whether Glenn Wilder thinks that the Harvard admission policy is 1. fair 2. unfair 3. neutral 4. I don't know the opinion of this person. We also ask each Turker to annotate Glen Wilder's emotions expressed in the post. We ask the Turker to do the same for Patrick O'Neil and Preston Moore. Figure 2 shows the distribution of sessions which includes different numbers of replies in each thread. The average number of replies in each thread in the dataset is 3.86, the median number of replies is 5.

In our dataset, the hotel reviews are highly focused and opinion rich with little irrelevant information, these reviews always talk about hotels or some aspects of a hotel, such as its location, cleanliness, service and price. Also, there is no interactions between reviewers, which means a reviewer cannot comment on another review.

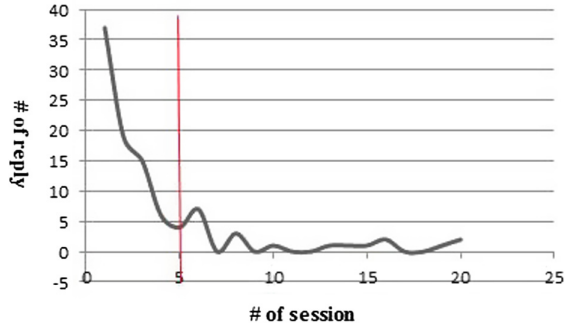


Fig. 2. Thread distributions in HAP.

Similar to the hotel reviews, the net neutrality dataset also does not contain any interactions between commenters. But unlike the hotel reviews which have clearly defined object-aspect relations between entities, the structure of the net neutrality comments is much more complex and there is no well-defined relations between the entities discussed in the comments (e.g., the policy itself, internet service providers, individual consumers, Netflix, pricing and innovation). Thus it can be very challenging to map different sentiments associated with different entities to an overall opinion about the net neutrality policy.

Comparing with the hotel reviews and net neutrality comments, the Twitter posts are much shorter - at most 140 characters. It involves a small number of interactions, such as retweet and reply. Since retweets normally do not change the sentiment and replies are relatively rare in our dataset, the impact of user interactions on Twitter sentiment analysis may not be as significant as that on HAP.

We did not perform additional annotations on the movie reviews since they are already annotated with sentiments and the dataset is mainly used to assess the appropriateness of employing supervised sentiment analysis for different domains.

3.2 Annotation Results

Since each data instance was annotated by three Turkers, we used the majority agreement as the ground truth labels. We also filtered out instances whose labels are “I do not know”. Table 3 displays the average agreement with the ground truth annotation for each domain. The results show that other than the HAP domain, the agreement with the ground truth opinion and emotion annotations from all the domains are high (near or above 90%). The most challenging case is HAP, because of the structural complexity of its posts, the agreement is only around 67% for the ground truth opinion annotation and 74% for the ground truth emotion annotation. Overall, for human annotators, other than the HAP domain, it is relatively easy for them to identify and agree on the emotions and opinions expressed in the posts.

Table 3. Agreement of annotated data.

	Opinion		Emotion	
	Majority Agreement	# of ground truth labels	Majority Agreement	# of ground truth labels
NN	0.91	458	0.86	431
Hotel	0.96	483	0.98	490
Tweet	0.912	899	0.95	949
HAP	0.669	84	0.74	101

3.3 Correlation Between the Opinion and Emotion Ground Truth

We have annotated two sets of ground truth, one is the emotions, the other is the opinions. Since opinions are more indirect and may require a deep understanding of the relationships between targets of emotions and a specific topic, we performed correlation analysis to assess their relations. We performed a Pearson chi-square test [15] to determine if the opinion and emotion ground truth are independent or correlated. If the p -value is smaller than 0.05, we can reject the null hypothesis of independence, which means there is a significant correlation between these two variables. Moreover, to measure the strength of this correlation, we calculated Cramer’s V . V may be viewed as the association between two variables as a percentage of their maximum possible variation. V can reach 1.0 when the two variables have equal marginals. A V value over 0.25, means the level of association is very strong. As shown in Table 4, other than net neutrality, the emotion and opinion ground truth are significantly correlated. However, based on the V values, the correlation on the HAP domain (0.357) is not as strong as those on the Hotel (0.98) and Twitter domain (0.93). Thus, for the net neutrality dataset, since the p -value is 0.24, we cannot reject the independence null hypothesis. Thus, it is possible that the two sets of ground truth are independent. This result indicates that if a SA tool is only capable of picking up emotion signals but not good at figuring out the relationships between targets of emotions and the subject of the opinion, it may not perform well on opinion mining on the NN or the HAP domain.

Table 4. Correlation between the opinion and the emotion ground truth.

Domain	p -value	Cramer’s V
NN	0.24	0.081
Hotel	<0.0001	0.98
Tweet	<0.0001	0.93
HAP	<0.0001	0.357

4 Empirical Study

To evaluate how different sentiment analysis tools perform on different datasets, we employed different tools. Among them, two are commercial state-of-the-art tools, two are open-source tools. Also, in terms of the learning methods, two of them use supervised sentiment classification and two of them use unsupervised sentiment analysis. All of them achieved over 75% prediction accuracy based on test data from their training domains.

4.1 Supervised Sentiment Analysis

Supervised methods consider sentiment classification as a standard classification problem in which labeled data are used to train a classifier. Many existing supervised sentiment analysis engines either provide pre-trained models or allow users to re-train their models using user-provided training data.

In our experiment, we used a commercial sentiment analyzer called TextProcessing which provides a pre-trained sentiment analysis model. The model was trained using annotated data from both the movie review domain and the Twitter domain. The movie review data come from [11] which are publicly available. It contains 1000 positive and 1000 negative reviews. The Twitter dataset is private and not available to us. Since TextProcessing is trained on two different domains, it is difficult for us to test the influence of domain difference on the analysis results. To overcome this, we also used a Naive Bayes-based text classifier to build a sentiment analyzer using the training examples from the movie review domain. To test the performance of our Naive Bayes sentiment analyzer, we randomly split the dataset into a training set (75%) and a testing set (25%). We repeat the process five times and the average prediction accuracy is 78%. The Naive Bayes sentiment analyzer used in the following experiments was trained using all 2000 annotated movie reviews. Because our training data have only two sentiment values: positive and negative; we only keep the positive and negative cases in our test data. Table 5 shows the statistics of the test datasets used to evaluate the performance of the two supervised SA tools on four different domains.

Table 5. Testing data of supervised tool.

Domain	Opinion		Emotion	
	Naive Bayes	TextProcessing	Naive Bayes	TextProcessing
NN	458	458	354	431
Hotel	483	483	474	490
Tweet	530	899	535	949
HAP	55	84	76	101

Figure 3 includes the prediction results based on the opinion ground truth. It shows that Naive Bayes analyzers performed the best on the hotel data.

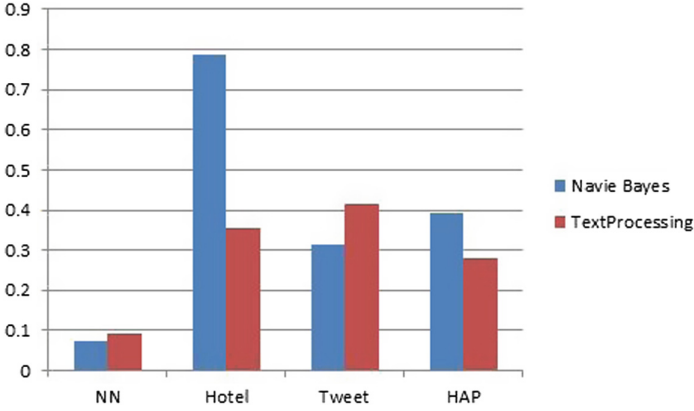


Fig. 3. Performance of supervised tools on opinion prediction.

The performance deteriorated significant on the HAP data. The Naive Bayes analyzer also performed significantly worse on the Twitter data. In contrast, the TextProcessing analyzer performed the best on the Twitter data. This may be due to the fact that a part of its training data came from Twitter. Surprisingly, both analyzers performed the worst on the Net Neutrality data since for humans, the HAP dataset is the most difficult one while the net neutrality data being relatively easy.

Figure 4 shows the evaluation results against the emotion ground truth. Again, the Naive Bayes classifier worked the best on the Hotel domain and the worst on the net neutrality domain. And the TextProcessing Analyzer worked the best on the HAP domain and the worst on the net neutrality domain.

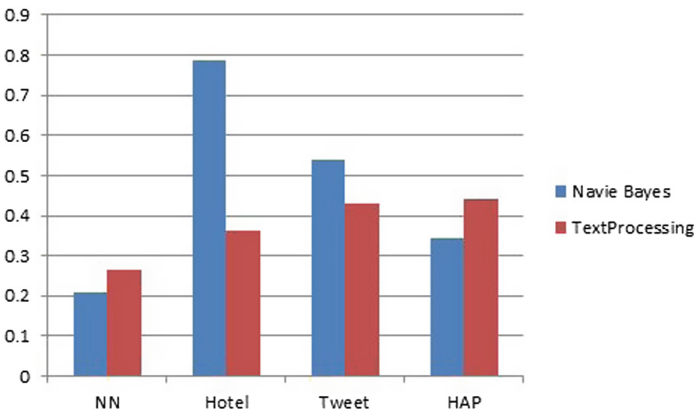


Fig. 4. Performance of supervised tools on emotion prediction.

When comparing their performance in predicting emotions and opinions, the TextProcessing analyzer performed much better in predicting emotions on NN and HAP domain. The Naive Bayes classifier captured emotion better on on NN and Twitter domain. The expression of emotion and opinion are very similar on the hotel, so the both analyzers performed similarly. Likewise, the TextProcessing worked similarly on the Twitter domain. It is worth noting that the Naive Bayers classifier performed better in prediction emotions on the Twitter domain. The Naive Bayers classifier trained with movie review data, the results shows the training set is more helpful to capture emotion signal instead of true opinion on different domain.

4.2 Unsupervised Sentiment Analysis

For unsupervised sentiment analysis, we employed Semantria, a commercial tool and SANN an open source sentiment analyzer. Both tools produce three sentiment labels: positive, negative and neutral.

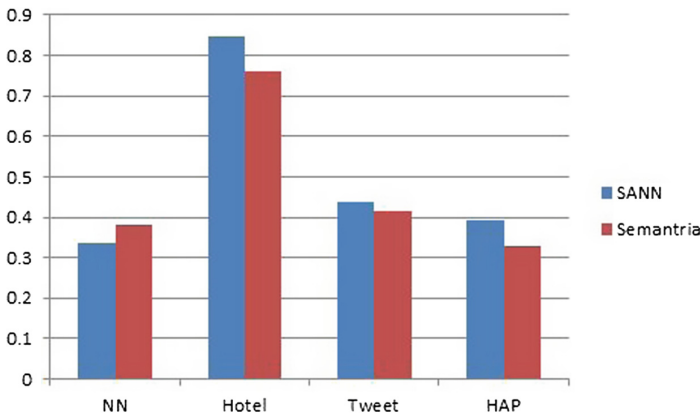


Fig. 5. Performance of unsupervised tools on opinion prediction.

Figure 5 shows the evaluation results against the opinion ground truth. The performance of SANN and Semantria are very similar - both of them achieved about 0.8 accuracy on the hotel data. Accuracy on tweet is both about 0.45. They performed the worst on the net neutrality and the HAP dataset with a prediction accuracy around 0.3.

Figure 6 shows the evaluation results against the emotion ground truth. Similarly, both SANN and Semantria performed the best on the hotel data. Twitter however seems to be the most challenging for both tools in emotion detection (accuracy is around 0.2).

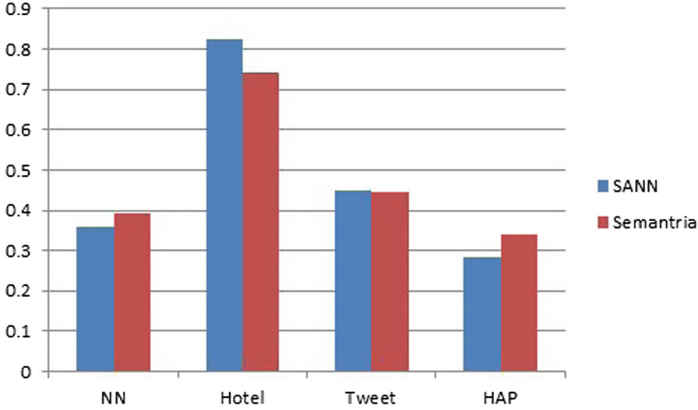


Fig. 6. Performance of unsupervised tools on emotion prediction.

4.3 Correlation Analysis Between Prediction Results and Two Sets of Ground Truth

We performed a Pearson chi-square test [15] to determine if two variables, the predicted value by a SA tool and the opinion/emotion ground truth, are correlated. As shown in Table 6, on the hotel dataset, since all the p-values are significantly less than 0.05 for all the tools for both opinion and emotion prediction, we can reject the null hypothesis and conclude that the predicted values are significantly correlated to both the opinion and emotion ground truth. To measure the strength of this correlation, we calculated Cramer’s V . As shown in Table 6, since all the V s on the hotel dataset are greater than 0.25, this indicates a strong correlation between the predicted values and the ground truth. Moreover, the two unsupervised tools SANN and Semantria performed well on the Twitter dataset for both opinion and emotion prediction. They also performed well on the net neutrality dataset for sentiment prediction. In contrast, none of the tools performed well in predicting opinions on the net neutrality dataset. Most of the tools also performed badly on the HAP dataset for both opinion and emotion prediction.

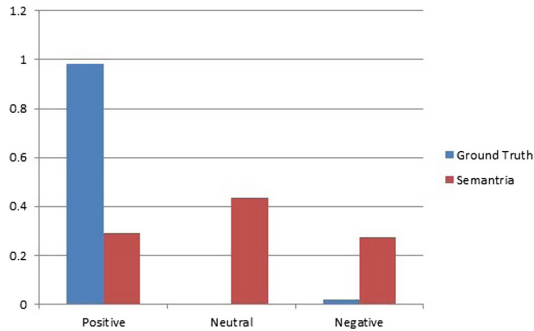
5 Domain Analysis

As we have shown in the previous section, domain differences have significant impact on sentiment analysis performance. If applied properly (e.g., to hotel reviews), the sentiment results may provide useful insight. If not careful and apply them mindlessly, the results can be meaningless or even misleading. For example, if we plot the sentiment analysis results from Semantria on the Net Neutrality dataset, we would believe that the public opinions towards net neutrality is ambivalent: 27% negative, 29% positive and 44% neutral (See Fig. 7). In fact

Table 6. Pearson chi-square test and Crammer’s V.

Method	Measure	NN		Hotel		Tweet		HAP	
		O	E	O	E	O	E	O	E
Navie Bayer	p -value	0.432	0.047	<0.0001	<0.0001	0.195	0.0985	0.21	0.618
	Crammer’s V	0.035	0.105	0.262	0.27	0.082	0.001	0.144	0.058
Text Processing	p -value	0.678	0.007	<0.0001	<0.0001	0.01	0.0002	0.82	0.25
	Crammer’s V	0.041	0.127	0.376	0.277	0.121	0.156	0.084	0.15
SANN	p -value	0.105	0.007	<0.0001	<0.0001	<0.0001	<0.0001	0.571	0.054
	Crammer’s V	0.095	0.385	0.503	0.357	0.213	0.23	0.117	0.196
Semantria	p -value	0.326	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.531	0.47
	Crammer’s V	0.067	0.233	0.587	0.409	0.166	0.213	0.122	0.122

Note: O is opinion, E is emotion

**Fig. 7.** Distribution of true opinion and Semantria’s results on net neutrality.

the real public opinion based on the ground truth annotation is un-ambiguously supportive: 97% support, 3% against and 0% neutral.

In the following, we investigate whether it is possible to automatically compute a set of effectiveness indicators to guide us in assessing the appropriateness of applying a sentiment analysis tool to a given dataset. For unsupervised methods, the effectiveness of a sentiment analysis tool is mainly determined by the properties of the target domain (e.g., complexity). For supervised methods, in addition to domain complexity, we hypothesize that the effectiveness can also be affected by the differences between the source and the target domain. In the following, we empirically verify the usefulness of several effectiveness indicators including *domain similarity*, *data genre*, *structure complexity* and *vocabulary complexity*.

5.1 Domain Similarity

For a supervised Sentiment Analysis tool, the similarity between the target and the source domain may have significant impact on sentiment analysis results. Among the two supervised tools, the pre-trained TextProcessing model was

trained on both movie reviews and Tweets while the Naive Bayes classifier was trained only on the movie review data. Since we don't have access to the Twitter training data used in TextProcessing, here we focus on the Naive Bayes Classifier. We computed two measures to assess the similarity: the *cosine similarity* and the χ^2 *similarity*. The *cosine similarity* is frequently used in information retrieval to measure the similarity between a search query and a document [18]. Here, we first construct two word vectors, one for all the movie reviews from the training data, one for all the text in a target domain (e.g., the hotel reviews). The length of a domain vector is the size of the entire vocabulary from all five domains. We then compute the cosine similarity between these two word vectors. We also computed the χ^2 *similarity* since it was shown to be the best one for assessing corpus similarity [8]:

$$\chi^2 = \sum \frac{(o - e)^2}{e}$$

Here, o is the observed frequency, e is the expected frequency. For each word, we calculate its occurrences in each corpus. If the size of corpus 1 and 2 are N_1 , N_2 , the word W has observed $O_{w,1}$ times in corpus 1 and $O_{w,2}$ times in corpus 2, then the expected frequency $e_w = \frac{N_1 * (O_{w,1} + O_{w,2})}{N_1 + N_2}$. When $N_1 = N_2$, the $e_w = \frac{O_{w,1} + O_{w,2}}{2}$. Since the χ^2 measure is not normalized, it does not permit direct comparison between corpora of different sizes [8]. As a result, for each domain, we constructed a new corpus, all with the same size by randomly sampling posts from each domain. In our experiment, the sample corpus size was set to be 25000 tokens. Based on our computation, the domain similarity ranks are:

$$HAP > HOTEL > NN > Tweet.$$

The most similar corpus to the movie corpus is HAP, while the Twitter corpus is the most different (Table 7).

Table 7. Corpus similarity between training dataset and testing dataset.

	$\cos(\theta)$	χ^2
NN	0.26	24000
Hotel	0.32	22427
Tweet	0.15	38034
HAP	0.45	21100

5.2 Genre

We also believe that the genre of text may impact the effectiveness of a sentiment analyzer. Here we categorize a text into three types: *review*, *comment* and *other*. Among them, reviews are often collected from dedicated review sites.

Each review contains explicit opinions about an obvious target. It has little irrelevant information. Also, there is a simple object-aspect relationship between the entities in a typical review (e.g., the screen of a digital camera). In our datasets, both the movie reviews and the hotel reviews belong to this category. Moreover, similar to reviews, comments are also opinion-rich. But the relationship between different entities in a comment is not well-defined. Also, due to the interactions between different commenters, correct sentiment analysis may require proper understanding of the conversation context, which makes comment-based sentiment analysis very challenging. In our datasets, both the FCC Net Neutrality dataset and the HAP dataset belong to this category. Finally, we categorize the Twitter data as *other* since they are collected based on keyword search and they can be almost anything. Simply speaking, the current sentiment analysis tools performed the best on reviews but poorly on comments or Tweets.

5.3 Structure Complexity

In sentiment analysis, complex domain often makes sentiment analysis difficult. Here, we first define a few measures on structure complexity. A straight-forward indicator of structure complexity is the average length of the posts in a domain. The ranking according to the length measure is:

$$\begin{aligned} Hotel &> NN > HAP > Tweet \\ 162.5 &> 68.39 > 58.84 > 15.78. \end{aligned}$$

Thus, hotel reviews tend to be much longer than the others. Due to the size constraints, Tweets are the shortest.

The second structure complexity indicator is the percentage of posts with external references. For example, in the following tweet: *Hillary Clinton: President Hopeful or Hopeless?* <http://wp.me/p3UNnh-BC>. Without opening the actual content using the URL, it is hard to know what the author's opinion is. The ranking according to the measure is:

$$\begin{aligned} Tweet &> HAP > NN > Hotel \\ 0.05 &> 0.001 > 0.0001 > 0. \end{aligned}$$

Thus, in these datasets, Tweets tend to have many embedded links while Hotel reviews are always self-contained without any external links.

The third structure complexity indicator is the average depth of a conversation thread, which is used to assess the complexity in user interactions. The ranking according to the average depth of a thread is:

$$HAP(4.8) > Tweet(1.37) > NN(1) = Hotel(1).$$

Based on this measure, HAP is the most complex domain with an average thread depth of about five. In contrast, both the NN and hotel reviews do not contain any user interacts.

Based on the performance of the tools, among the three structure complexity indicators, the post length seems to have little impact on the prediction accuracy while external links and tread depth can make opinion analysis more difficult.

5.4 Vocabulary Complexity

Entropy is a measurement of vocabulary’s homogeneity. Given a sequence of words i.e. words $(w_1, w_2, w_3, \dots, w_n)$, the entropy can be computed using:

$$H = - \sum_{W_i^n \in L} P(W_i) * \log P(W_i)$$

To normalize it, we calculated the relative entropy $H_{rel} = \frac{H}{H_{max}}$, where H_{max} is the max entropy which occurs when all the words have a uniform distribution, thus $p = 1/||w||$. To avoid the impact of corpus size, we construct four new corpora with equal size, each by randomly sampling posts from each of the four original corpora. As shown in Fig. 8, computed relative entropy is no longer sensitive to corpus size. When we varied the sample corpus size from 1000 to 25000, there is no significant difference in computed relative entropy.

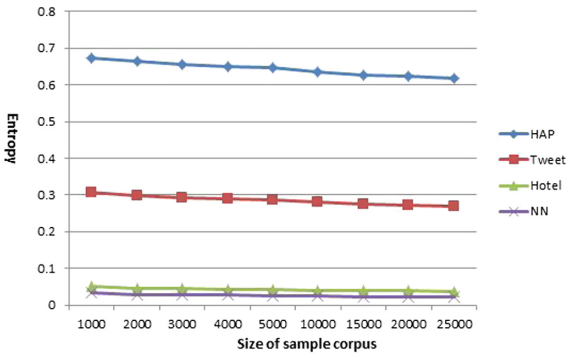


Fig. 8. Entropy of each corpus.

As shown in Fig. 8, the vocabulary complexity of HAP is much higher than the other three. It is also pretty high for the twitter domain. The values of hotel and NN are very close, both have low entropy. This is an indication that their vocabularies are relatively homogeneous.

5.5 Result Analysis

Based on our results, HAP should be the most difficult domain for sentiment analysis. Its genre is *comment*, one of the more complex genres for sentiment analysis. Its vocabulary complexity based on relative entropy is the highest. In terms of average thread depth, its structure complexity is the highest as well. This has been proven to be true for both humans (based on the ground truth annotation) and for computers (The prediction accuracy is about 0.3 for all the supervised and unsupervised tools we tested). In contrast, the hotel review domain should be relatively easy for sentiment analysis. Its genre is review, one of

the easiest. It has little or no external references and user interactions. Moreover, its vocabulary complexity is one of the lowest, which makes it an ideal domain for sentiment analysis.

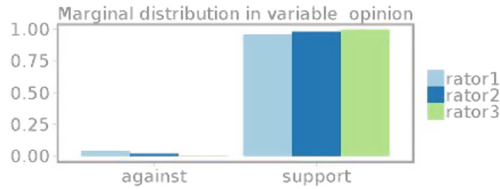


Fig. 9. Annotated opinion distribution on NN.

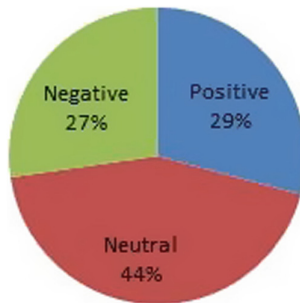


Fig. 10. Detected result distribution on NN.

It is worth noting that our sentiment analyzers performed poorly on NN. Based on our domain analysis, its vocabulary complexity is among the easiest (very close to the hotel domain), its average post length is much shorter than hotel reviews. It also does not have many external references and user interactions. It is a surprise to see that all the tools performed poorly on this dataset. By inspecting the ground truth data, we found that it is highly unbalanced. As shown in Fig. 9, over 95% people support net neutrality. In contrast, the output from Semantria has a very different distribution of sentiment (see Fig. 10). After inspecting the positive and negative comments predicted by Semantria, we found that the system is unable to map the sentiment expressed in the text to a opinions toward net neutrality since the relationships between them are very complex. For example, a person may express “Net Neutrality is great for innovation” or “Comcast is very greedy”. Although the sentiment in the first message is “positive” while the second one is negative, the authors of both comments support net neutrality. To get it right, sophisticated inferences of the relationship between Comcast and net neutrality is needed. So far, most of the sentiment analysis tools are not capable of handling this type of inference.

6 Conclusion

Sentiment analysis has been used frequently by businesses, organizations and individuals to assess public opinions and gain insights. In this paper, we empirically analyze the appropriateness of applying sentiment analysis tools in five different domains. Our results demonstrated the importance of understanding the potential pitfalls associated with applying these tools in a given domain. We also proposed several *effectiveness indicators* which can be computed automatically to signal potential problems.

In our current study, we only compare datasets vertically which means all of them are from different data sources. In the future, we want to compare the domain horizontally, collecting data on different topics from the same source (e.g., on Twitter). We also noticed the importance in understanding the relationships between different entities in a domain and the target opinion. We plan to develop new measures that can capture the complexity of entity-opinion relationships in a domain.

References

1. Andreevskaia, A., Bergler, S.: Unsupervised sentiment analysis with emotional signals. In: 11th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2006 (2006)
2. Belkin, D.: Harvard accused of bias against Asian-Americans. Wall Street J. (2015). <http://www.wsj.com/articles/asian-american-organizations-seek-federal-probe-of-harvard-admission-policies-1431719348>
3. Bob Lannon, A.P.: What can we learn from 800,000 public comments on the FCC's net neutrality plan? @ONLINE (2014)
4. Chung, W., Zeng, D.: Social-media-based public policy informatics: sentiment and network analyses of U.S. immigration and border security. J. Assoc. Inf. Sci. Technol. **67**(7), 1588–1606 (2015)
5. Diakopoulos, N.A., Shamma, D.A.: Characterizing debate performance via aggregated twitter sentiment. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2010, pp. 1195–1198. ACM, New York (2010)
6. Ghose, A., Ipeirotis, P.G., Sundararajan, A.: Opinion mining using econometrics: a case study on reputation systems. In: Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (2007)
7. Hu, X., Tang, J., Gao, H., Liu, H.: Unsupervised sentiment analysis with emotional signals. In Proceedings of the 22nd International Conference on World Wide Web, WWW 2013, Republic and Canton of Geneva, Switzerland, pp. 607–618. International World Wide Web Conferences Steering Committee (2013)
8. Kilgariff, A., Rose, T.: Measures for corpus similarity and homogeneity. In: 3rd Conference on Empirical Methods in Natural Language Processing (1998)
9. Liu, B.: Sentiment Analysis and Opinion Mining. Morgan and Claypool, New York (2012)
10. Maynard, D., Funk, A.: Automatic detection of political opinions in tweets. In: García-Castro, R., Fensel, D., Antoniou, G. (eds.) ESWC 2011. LNCS, vol. 7117, pp. 88–99. Springer, Heidelberg (2012). doi:[10.1007/978-3-642-25953-1_8](https://doi.org/10.1007/978-3-642-25953-1_8)

11. Pang, B., Lee, L.: A sentimental education: sentiment analysis using subjectivity. In: Proceedings of ACL, pp. 271–278 (2004)
12. Pang, B., Lee, L.: Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* **2**(1–2), 1–135 (2008)
13. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, Stroudsburg, PA, USA, vol. 10, pp. 79–86. Association for Computational Linguistics (2002)
14. Pappas, N., Katsimpras, G., Stamatatos, E.: Distinguishing the popularity between topics: a system for up-to-date opinion retrieval and mining in the web. In: 14th International Conference on Intelligent Text Processing and Computational Linguistics (2013)
15. Plackett, R.L.: Karl pearson and the chi-squared test. *Int. Stat. Rev.* **51**, 59–72 (1983)
16. Read, J., Carroll, J.: Weakly supervised techniques for domain-independent sentiment classification. In: Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion, TSA 2009, New York, NY, USA, pp. 45–52 (2009)
17. Shi, H.-X. and Li, X.-J.: A sentiment analysis model for hotel reviews based on supervised learning. In: 2011 International Conference on Machine Learning and Cybernetics (ICMLC), vol. 3, pp. 950–954 (2011)
18. Singhal, A.: Modern information retrieval: a brief overview. *IEEE Data Eng. Bull.* **24**(4), 35–43 (2001)
19. Turney, P.D.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL 2002, Stroudsburg, PA, USA, pp. 417–424. Association for Computational Linguistics (2002)
20. Wang, H., Can, D., Kazemzadeh, A., Bar, F., Narayanan, S.: A system for real-time twitter sentiment analysis of 2012 U.S. presidential election cycle. In: Proceedings of the ACL 2012 System Demonstrations, ACL 2012, Stroudsburg, PA, USA, pp. 115–120. Association for Computational Linguistics (2012)
21. Wang, H., Lu, Y., Zhai, C.: Latent aspect rating analysis without aspect keyword supervision. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2011, New York, NY, USA, pp. 618–626 (2011)
22. Peng, W., Park, D.H.: Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. In: The International AAAI Conference on Web and Social Media, ICWSM (2011)
23. Zhou, X., Tao, X., Yong, J., Yang, Z.: Sentiment analysis on tweets for social events. In: 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD), pp. 557–562 (2013)