An Empirical Study of the Effectiveness of using Sentiment Analysis Tools for Opinion Mining

Tao Ding and Shimei Pan
Department of Information Systems, University of Maryland Baltimore County, 1000, Hilltop Cir, Baltimore, U.S.A.

Keywords: Content Analysis, Sentiment Analysis, Performance Measure.

Abstract: Sentiment analysis is increasingly used as a tool to gauge people's opinions on the internet. For example, sentiment analysis has been widely used in assessing people's opinions on hotels, products (e.g., books and consumer electronics), public policies, and political candidates. 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 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. In this paper, we present an empirical analysis of the effectiveness of using existing sentiment analysis tools in assessing people's opinions in five different domains. We also proposed several *effectiveness indicators* that can be computed automatically to help avoid the potential pitfalls in misusing a sentiment analysis tool.

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 us gain insight and make informed decisions. For example, it can help us determine if a marketing initiative is driving the planned responses, or determine whether consumers like a new product just launched or not, or people's reaction to a political debate (Diakopoulos and Shamma, 2010; Wang et al., 2012). 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 try to 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 (Ghose et al., 2007), services (Shi and Li, 2011), social events (Zhou et al., 2013), political events (Diakopoulos and Shamma, 2010), political candidates, and public policies (Wang et al., 2012; Chung and Zeng, 2015).

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 applications 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'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 people determine whether a SA tool is appropriate for one's data, we

Ding, T. and Pan, S.
An Empirical Study of the Effectiveness of using Sentiment Analysis Tools for Opinion Mining.
In *Proceedings of the 12th International Conference on Web Information Systems and Technologies (WEBIST 2016) - Volume 2*, pages 53-62
ISBN: 978-989-758-186-1

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 and comprehensive 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 using sentiment analysis in assessing public opinions. Our results can also help raise the awareness of the potential pitfalls associated with the misuse of sentiment analysis.
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.

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 frequently, in a broad sense is defined as the computational study of opinions, sentiments and emotions expressed in text (Pang and Lee, 2008). According to (Liu, 2012), 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 are 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 approaches (Maynard and Funk, 2012). 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 (Pang et al., 2002). 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 API[1], others require users to provide labeled data and then train their own prediction models, such as Google Prediction API[2], NLTK text classification API[3].

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 (Andreevskaia and Bergler, 2006; Wei Peng, 2011). (Turney, 2002) 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 (Read and Carroll, 2009) proposed a weakly-supervised technique, using a large collection of unlabeled text to determine sentiment. They used PMI (Turney, 2002), semantic spaces, and distributional similarity to measure similarity between words and polarity prototype. The results were less dependent on the domain, topic and time-period represented by the testing data. In addition, Hu (Hu et al., 2013) investigated whether models of emotion signals can potentially help sentiment analysis.

[1] http://text-processing.com/docs/sentiment.html

[2] https://cloud.google.com/prediction/docs

[3] http://www.nltk.org/api/nltk.classify.html

So far, hundreds of commercial state-of-the-art tools available for automatic sentiment analysis, such as Semantria[4], SentimentAnalyzer[5], SentiStrength[6], MLAnalyzer[7], TextProcessing[8]. 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 (Pang and Lee, 2004b). 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 SANN [9] (Pappas et al.,).

Table 1: Selected tools.

Method	Tool
Supervised	Naive Bayer TextProcessing
Unsupervised	SANN Semantria

3 DATA COLLECTION

To evaluate the impact of domain differences on sentiment analysis, we included five datasets (Table 2):

1. **Hotel Reviews (Hotel):** The dataset was originally used in (Wang et al., 2011). 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.
2. **Net Neutrality(NN):** The US Federal Communications Commission (FCC) (Bob Lannon, 2014) 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 14-28 [10]). 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 Admission Policy (HAP):** Recently, *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 (Belkin, 2015). 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 senti-

[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

[10] https://www.fcc.gov/rulemaking/most-active-proceedings

ment analysis tool, such as TextProcessing, used these polarity data to train sentiment classifier. We compare other four domains with movie domain in experiments regarding performance of supervised tools.

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 obtain a ground truth about the true opinion expressed in the text, we made use of Amazon's Mechanical Turk(AMT) to annotate the overall opinion expressed in each review, comment and tweet. 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 opinion expressed in the text. To ensure the quality of the ground truth data, each post is annotated by three different annotators. All the annotators also have to be qualified based on the following criteria: they must have submitted over 5000 tasks with an acceptance rate of over 95%.

Specifically,

For hotel review, each participant is asked whether the author 1. likes the hotel; 2. dislikes the hotel; 3. is neutral; 4. does not know the author's opinion.

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 comments about net neutrality, we asked each turker whether the author : 1. supports net neutrality; 2.is against net neutrality; 3. is neutral; 4. does not know the author's opinion. Five hundreds hotel reviews and 500 comments on NN are selected randomly to annotate. Here is an example comment from the 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.

For Twitter posts, we asked each turker to rate whether the author 1. supports Hillary Clinton 2. does not support Hillary Clinton, 3. is neutral 4. I do not know the opinion of the author. We randomly selected 1000 out of 7237 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 in these comments. 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 the correct opinion, instead of providing a comment for annotation, 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.

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

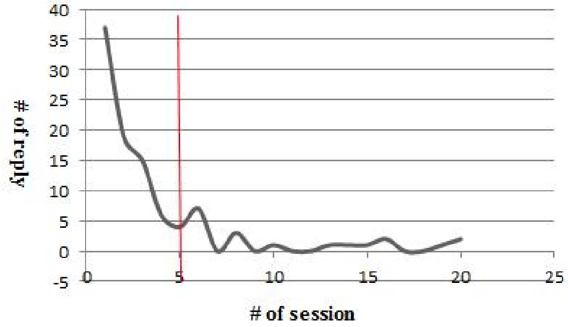

Figure 1: Thread Distribution in HAP.

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 reviewer's comment.

Similar to the hotel reviews, the net neutrality dataset also does not contain any interactions between commenters. But unlike the hotel reviews which has clearly defined object-aspect relations between entities, the net neutrality topic 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 users, netflix, pricing and innovation). Thus it can be very challenging to map different sentiments associated with different entities to an overall opinion towards the net neutrality policy.

Comparing the hotel reviews and net neutrality comments, the Twitter posts are much shorter - at most 140 characters. It involved 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 sentiment analysis on Twitter may not be as significant as that on HAP.

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" based on majority agreement. Table 3 displays the average agreement with the majority-based ground truth annotation for each domain. The results show that other than the HAP domain, the agreement from all the three other domains are high (above 90%). The most challenge case is HAP, because of the complex structure, the agreement with the ground truth is only around 67% for human annotators.

Table 3: Majority Agreement of annotated data.

	Majority Agreement	# of ground truth label
NN	0.91	431
Hotel	0.96	483
Tweet	0.912	899
HAP	0.669	84

4 EMPIRICAL STUDY

To evaluate how different sentiment analysis tools perform on different datasets, we employed four different tools. Among them, two are commercial state-of-the-art tools, two are open-source tools. Also, in terms of the learning methods employed, 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 the same domain.

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 (Pang and Lee, 2004a) 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 each domain 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 on all 2000 annotated movie reviews. Because our training data have only two sentiment values: positive and negative; we filtered the *neutral* category from our test data showed in column 1 from table 4.

Table 4: Testing data of supervised tool.

Domain	Naive Bayes	TextProcessing
NN	354	431
Hotel	472	483
Tweet	530	899
HAP	55	84

From our results shown in figure 2, both analyzers performed the best on the hotel data. Their performance deteriorated significant on the HAP data. The Naive Bayes analyzer also performed significantly worse on the Twitter data. In contrast, the TextProcessing analyzer didn't deteriorate as much. 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.

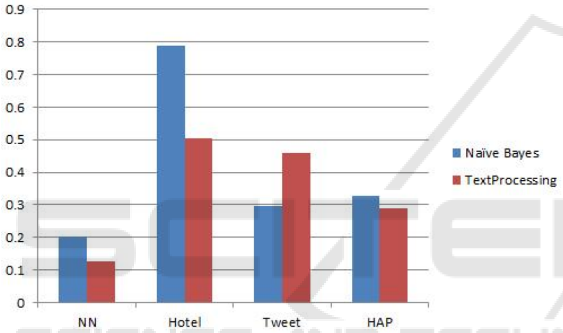

Figure 2: Performance of supervised tools.

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.

The performance of SANN and Semantria are very similar - both of them achieved 0.8 accuracy on the hotel data. Accuracy on tweet is both 0.45. They performed worst on both the net neutrality and the HAP data a with prediction accuracy around 0.3.

4.3 Correlation between the Predicted Sentiment and Ground Truth

We performed a Pearson chi-square test (Plackett, 1983) to determine if two variables, the ground truth and the predicted sentiment analysis results, are independent. If the *p*-value is smaller than 0.05, we can reject the null hypothesis of independence. Since the *p*-value on NN and HAP are greater than 0.05, we can not reject the null hypothesis. Thus, it is possible that the predicted sentiment and the ground truth are not related. Since the *p*-value on the hotel and the Twitter data are mostly less than 0.05, we can reject the null hypothesis of independence and conclude there is a significant correlation between the predicted value and the ground truth. To measure the strength of relationship between the predicted results and the ground truth, we calculated Crammer'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 only when the two variables have equal marginals. If the V value is over 0.25, this means the level of association is very strong. Table 5 shows that all four tools performed well on the hotel reviews. On Twitter data, the predicted results by TextProcessing, SANN, and Semantria are strongly correlated with the ground truth. HAP's V value is around 0.2, which shows a moderate correlation. There is no or negligible relationship between the predicted sentiment and the ground truth on the NN dataset.

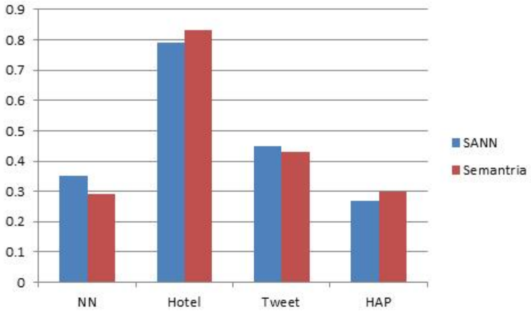

Figure 3: Performance of unsupervised tools.

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 Figure 4). In fact the real public opinion based on the ground truth annotation is unambiguously 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 appropriate-

Table 5: Pearson chi-square test and Crammer's V.

Domain	Navie Bayer		Text-Processing		SANN		Semantria	
	p-value	Crammer's *V*	*p*-value	Crammer's *V*	*p*-value	Crammer's *V*	*p*-value	Crammer's *V*
NN	0.07	0.102	0.69	0.052	0.14	0.09	1	0.01
Hotel	0.0004	0.84	0.0004	0.9	0.0004	0.937	0.0004	0.97
Tweet	0.07	0.076	0.00001	0.27	0.0004	0.27	0.0009	0.27
HAP	0.24	0.18	0.1	0.24	0.06	0.25	0.6	0.18

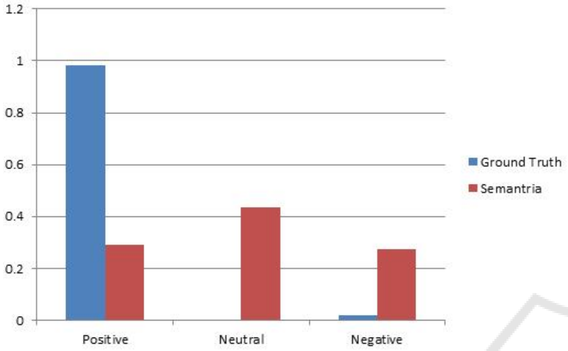

Figure 4: Distribution of Ground truth and Semantria's results on Net Neutrality.

ness 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 additional 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

Domain similarity may have significant impact on sentiment analysis results. In our experiment on evaluating 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. To measure the similarity between each target and training domain, we computed two measures to assess their 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 (Singhal, 2001). 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 domain). The length of each 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 (Kilgarriff and Rose, 1998):

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

o is observed frequency, e is expected frequency. For each of n words, we calculate the number of occurrences in each corpus. If the size of corpus 1 and 2 are N_1, N_2, the word W has observed frequencies $O_{w,1}$ in corpus 1, $O_{w,2}$ in corpus 2, then 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 (Kilgarriff and Rose, 1998). As a result, for each domain, we constructed a new corpus with the same size by randomly sampling posts from each domain. In our experiment, the sample corpus size was set to 25000 tokens. The ranking of similarity is:

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

The most similar corpus to the movie corpus is HAP, while the Twitter corpus is the most different.

Table 6: 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 single or multiple comments are not well-defined. Also, due to the interactions between different commenters, correct sentiment analysis may require proper understanding of conversation context, which makes comment-based sentiment analysis very challenging. In our datasets, both the FCC Net Neutrality data and the HAP data 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, sentiment analysis performs the best on reviews but poorly on comments.

5.3 Structure Complexity

In sentiment analysis, complex domain often makes sentiment analysis more difficult. Here, we first define a few measures on structure complexity. Later we also propose a measure for vocabulary 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:

$$Hotel > NN > HAP > Tweet$$

$$162.5 > 68.39 > 58.84 > 15.78.$$

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 open content using the URL , it is hard to know what the author's opinion is. The ranking according to the measure is:

$$Tweet > HAP > NN > Hotel$$

$$0.05 > 0.001 > 0.0001 > (\ll 0.0001).$$

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

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

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.

5.4 Vocabulary Complexity

Entropy is a measurement of vocabulary's homogeneity. Given a sequence of words i.e. words$(w_i, w_2, w_3..., w_n)$, 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 Figure 5, relative entropy is not sensitive to corpus size. When we varied the sample corpus size from 1000 to 25000, there is no significant difference in their relative entropy.

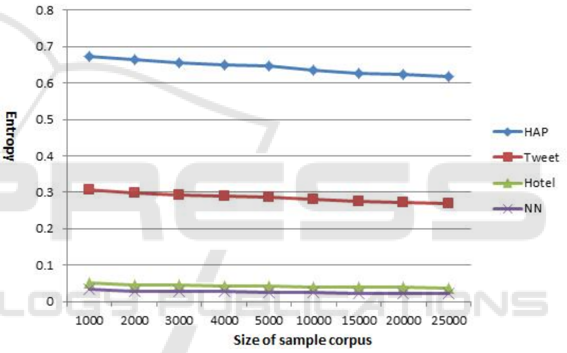

Figure 5: Entropy of Each Corpus.

Based on this measure, the vocabulary complexity of HAP is much higher than the other three. 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 data) 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.

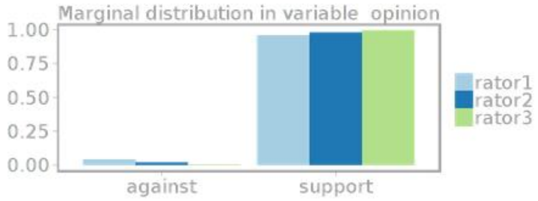

Figure 6: Annotated opinion distribution on NN.

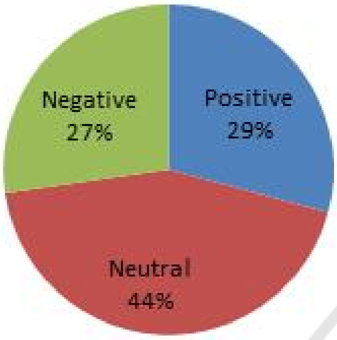

Figure 7: Detected Result Distribution on NN.

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 (see Figure 6): over 95% people support net neutrality. In contrast, the output from Semantria has a very different distribution of sentiment (see Figure 7). 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 person's opinion 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 becomes increasingly popular for businesses, organizations and individuals to assess public opinions and gain insight. In this study, we empirically investigate the effectiveness of different sentiment analysis tools on different domains. Our results demonstrated the importance of the appropriate use of sentiment analysis tools and the potential pitfalls associated with using these tools mindlessly. We also proposed several *effectiveness indicators* which can be computed automatically to guide us to use them appropriately for opinion mining.

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 from the same source (e.g., on Twitter) but on different topics. We also noticed the importance in understanding the relationships between different entities and the target. We plan to find new measures that can capture the complexity of relationships in a domain.

REFERENCES

Andreevskaia, A. and Bergler, S. (2006). Unsupervised sentiment analysis with emotional signals. In *11rd Conference of the European Chapter of the Association for Computational Linguistics*, EACL-06.

Belkin, D. (2015). Harvard accused of bias against asian-americans. *The Wall Street Journal.*

Bob Lannon, A. P. (2014). What can we learn from 800,000 public comments on the fcc's net neutrality plan? @ONLINE.

Chung, W. and Zeng, D. (2015). Social-media-based public policy informatics: Sentiment and network analyses of u.s. immigration and border security. *Journal of the Association for Information Science and Technology*, pages n/a–n/a.

Diakopoulos, N. A. and Shamma, D. A. (2010). Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1195–1198, New York, NY, USA. ACM.

Ghose, A., Ipeirotis, P. G., and Sundararajan, A. (2007). Opinion mining using econometrics: A case study on reputation systems. In *In Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL*.

Hu, X., Tang, J., Gao, H., and Liu, H. (2013). Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 607–618, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Kilgarriff, A. and Rose, T. (1998). Measures for corpus similarity and homogeneity. In *3rd Conf. on Empirical Methods in Natural Language Processing*.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan and Claypool.

Maynard, D. and Funk, A. (2012). Automatic detection of political opinions in tweets. In *Proceedings of the 8th International Conference on The Semantic Web*, ESWC'11, pages 88–99, Berlin, Heidelberg. Springer-Verlag.

Pang, B. and Lee, L. (2004a). A sentimental education: Sentiment analysis using subjectivity. In *Proceedings of ACL*, pages 271–278.

Pang, B. and Lee, L. (2004b). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pappas, N., Katsimpras, G., and 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*.

Plackett, R. L. (1983). Karl pearson and the chi-squared test. *International Statistical Review*, 51:59–72.

Read, J. and Carroll, J. (2009). Weakly supervised techniques for domain-independent sentiment classification. In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion*, TSA '09, pages 45–52, New York, NY, USA. ACM.

Shi, H.-X. and Li, X.-J. (2011). A sentiment analysis model for hotel reviews based on supervised learning. In *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, volume 3, pages 950–954.

Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43.

Turney, P. D. (2002). 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 '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., and Narayanan, S. (2012). A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, pages 115–120, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wang, H., Lu, Y., and Zhai, C. (2011). Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 618–626, New York, NY, USA. ACM.

Wei Peng, D. H. P. (2011). 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.

Zhou, X., Tao, X., Yong, J., and Yang, Z. (2013). Sentiment analysis on tweets for social events. In *Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on*, pages 557–562.