Opinion Mining for Twitter

Rutu Mulkar-Mehta

Abstract

This paper is a summarization of the work I did for my internship at PriceWaterhouseCoopers.

Introduction

What is twitter?

Twitter is another social networking tool like Facebook, MySpace, orkut. Etc. It was launched in 2006 but gained popularity in 2007. It is a very simple and lightweight tool based on people you are following, and people following you. The status messages or "tweets" can only be 140 characters and are shared with followers. All the data is publically available and might appear in twitter search (unless otherwise specified by the user).

Who uses twitter?

Twitter is being used by individuals, businesses, radio stations, libraries, offices, performers, artists etc. It has become a broadcasting media where you can follow people/companies that you are interested in, and you will get all the information about them firsthand. Companies post new coupons and advertisements about their products directly to the consumers, your favourite artists post dates and venues of their latest performances, libraries and other offices mention their working hours and anything else that is happening out of the ordinary.

Impact of Twitter

Twitter has had an impact on us. Lately, news agencies are getting their news topics from trending topics of twitter. News agencies have started displaying "tweets" about the trending topic alongwith their story, to show direct opinions from real people. Some examples of the importance of twitter was seen during the earthquake in China (May 12th 2008), earthquake in Mexico (May 22nd 2009), the Iran elections (2009) etc.

Studies on the impact of Twitter (Krishnamurthy, Gill, and Arlitt 2008) talks about something (Java et al. 2007) talks about why twittering has become such a popular phenomenon. (Jansen et al. 2009) talks about how microblogging has become a replacement for the word of mouth sentiments. (Huberman 2009) talks about social attention of people, and how microblogs are making a bigger impact on people. (Leskovec et al. 2008) talks about the evolution of micro-blogs, and its impacts on society.

Sentiment Analysis

Previous Work

There has been a lot of work on sentiment analysis. (Hu and Liu 2004) talks about sentiment analysis on product reviews, and later elaborates his work in (Ding, Liu, and Yu 2008). (Zhuang, Jing, and Zhu 2006) and (Zhuang et al. 2006) talk about analysing sentiment from movie reviews. Other work includes (Kim and Hovy 2006), (Pang, Lee, and Vaithyanathan 2002), (Du and Tan 2009), (Wilson, Wiebe, and Hoffmann 2005) and (Glance et al. 2005).
(Greene and Resnik 2009) talks about extracting impicit sentiment in text. (Esuli and Sebastiani 2006) have worked on classifying sentiment words in Wordnet and (Angela 2008) has worked on classifying the polarity of adjetives.

The challange

All the previous work worked with structured review data with longer sets of sentences. E.g. "i recently purchased the canon powershot g3 and am **extremely satisfied** with the purchase . The camera is **very easy to use** , in fact on a recent trip this past week i was asked to take a picture of a vacationing elderly group . after i took their picture with their camera , they offered to take a picture of us . i just told them , press halfway , wait for the box to turn green and press the rest of the way . they fired away and the **picture turned out quite nicely**. (as all of my pictures have thusfar).

Our challenge was to work with raw data that might not be review data, and the length of each indiviual data unit was less than 240 characters. E.g. “iPhone audio jack went funny. Got worried. Turned out to be a little ball of fluff. Took it out with a toothpick. Everything fine now. Phew. Bank Lets Customers Deposit Checks by Taking Pics with an iPhone http://bit.ly/O9xAK”

Our Goal was to perform Sentiment Analysis on twitter corpus, and get overall sentiment about a specific product.

The business motivation for this is that it could be sold as a service to PwC clients to know the brand impact of their product within the first few hours of the release.

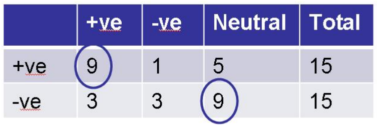
Twitter

The Twitter Database

Twitter has the advantage of having an unlimited supply of data about a trending topic. The twitter search API is very mature and can extract all tweets related to a specific search. The only limitation we have is the rate and data limitations imposed by twitter for search. Search can be performed back to one month, or 1500 tweets only (whichever is less). As a result we had to use a third part software to save our old searches and work with them. We used **The archivist** (arc) for this purpose.

Twitter Message Characteristics

- **Length:** 140 characters long
- **Creativity :** “If the iPhone were a Japanese movie, it would be Godzilla. The blackberry would be the screaming townspeople.”
- **Sarcastic:** “Waited 15 min for bus then firetrucks blocked intersection so had to walk, now dropped iPhone & cracked the face. Great start to the day.”
- **Multiple sentiments in a single tweet:** “love an iPhone but it is too expensive”
- **Links:** “Interesting: Blackberrys 26 Advantages over iPhone http://seekingalpha.com/a/3blz ”
- **Spam :**“Anybody want to win an iPhone 3GS? Follow @iphonecontests for all the best contest news for the iPhone and iPod Touch Please RT”
- **Twitter Vocabulary:**
  - *@username :* Reply to a tweet
  - *hashtags :* used to make the tweet searchable
    I love my #iPhone.
    I love my iPhone #iPhone.
  - *RT:* ReTweet
  - *Acronyms and Slang:*
    FTW: For The Win
    ROTLF: Rolling on the floor laughing



	+ve	-ve	Neutral	Total
+ve	9	1	5	15
-ve	3	3	9	15

Figure 1: Evaluation of the current twitter sentiment analysis

Previous Work

Twitter acquired the company Summize in July 2008. Summize performed some sentiment analysis on twitter data, and twitter claims to have included those sophisticated sentiment analysis techniques in twitter search. I tried to evaluate their current methods and ran a simple query for extracting positive and negative sentiment about iPhone. The search returned 15 positive and 15 negative tweets about the iPhone. The evaluation showed that out of the 15 positive tweets, 9 of the tweets were positive and the rest were either negative or neutral. For the negative tweets however, 9 out of 15 of the tweets were actually neutral. This showed that emoticons do not play a very important role in sentiment detection.

Dell Emoticon Experiment

I ran the emoticons experiment in a different domain for a larger corpus. The size of the corpus was 6554 tweets and the domain selected was tweets about the company Dell. We Extracted all tweets which contained :- in them. The system returned 38 tweets containing 6 different emoticons. They collectively conveyed 3 different sentiment types: positive (+), negative (-) and neutral ( ).
Examples:

```
(+) i wish he got me a dell inspiron :-(
```

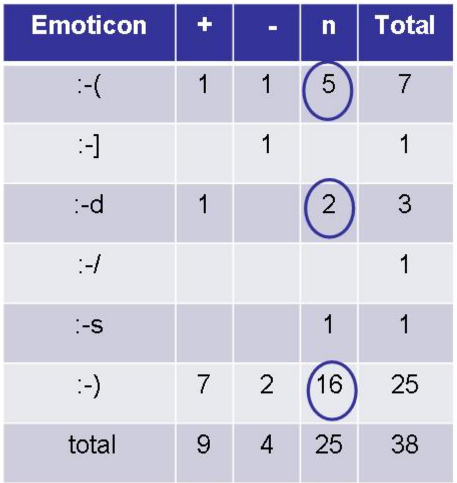
This is actually a positive emotion as the sadness is because of the lack of the dell. Another example of a misclassified negative sentiment is as follows:

```
(-) turning my dell netbook into
a mac netbook :-) in atlanta, ga
```

Other work on Twitter

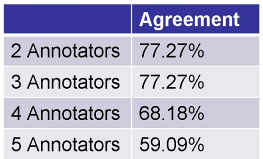
Apart from Summize there were 3 projects in the Natural Language class at Stanford University in 2009. Some analysis of these projects is shown below.

- **Project 3:** Here they collected their by the presence of smiles :) or frowns :(emoticons. As we already showed, emoticons are not a good measure for determining sentiment and hence their results did not seem very convincing.
- **Project 19:** Worked with general sentiment analysis, rather than sentiment about a single chosen topic.
- **Project 22:** Not enough information



Emoticon	+	-	n	Total
:-(	1	1	5	7
:-]		1		1
:-d	1		2	3
:-/				1
:-s			1	1
:-)	7	2	16	25
total	9	4	25	38

Figure 2: Evaluation of the presence of Emoticons in Tweets



	Agreement
2 Annotators	77.27%
3 Annotators	77.27%
4 Annotators	68.18%
5 Annotators	59.09%

Figure 3: Inter-Annotator agreement

Experiments Performed

Annotation Effort

- **What we want to annotate:** Mark tweets to be positive, negative or neutral
- **Annotation Corpus :** A set of 22 random tweets were selected
- **Annotators :** 5 annotators (Interns at CAR)

Bootstrapping using tweets

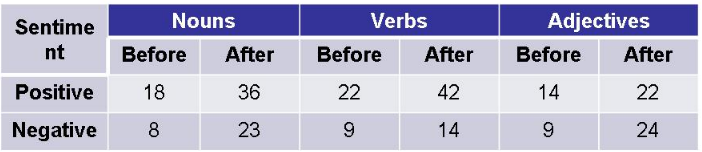
The following were the two subroutines we used for bootstrapping.

```
SUBROUTINE A
Start with seed words: love, hate
Extract all sentences with seed words
Filter out sentences where the
distance between seed word and theme word >7
Find the pattern of words between
theme word and seed word
Regular expression:
love (.*) iPhone
```

The examples of results were: My, to win my, tweeting from my, to win an, to win the, the new etc.

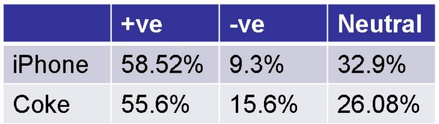
```
SUBROUTINE B
Start with seed words from subroutine A
and collect tweets
Regular Expression
(.*) to win an iPhone
```

The examples of results were: i'd like, Trying, i want, Hoping, would like etc.



Sentiment	Nouns		Verbs		Adjectives	
	Before	After	Before	After	Before	After
Positive	18	36	22	42	14	22
Negative	8	23	9	14	9	24

Figure 4: Results of bootstrapping using wordnet



	+ve	-ve	Neutral
iPhone	58.52%	9.3%	32.9%
Coke	55.6%	15.6%	26.08%

Figure 5: Precision of extracting positive sentiment tweets

Bootstrapping using Wordnet

WordNet is a large lexical database of English, Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms

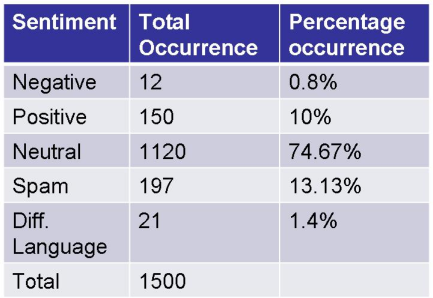
We annotated 300 sentences from the iPhone domain and collected a number of positive and negative words. Examples: (N)woes, issues, (V)hate, love , (A)unreliable, amazing etc. The following applications use this technology
twitrratr.com
twittersentiment.appspot.com
twentz.com

Experiment 1: Simple Keyword Search

This experiment checked for the presence or absence of a positive or negative keyword in the tweet. The words used were extracted from the wordnet bootstrapping experiment. We achieved a precision of 55Reason for high accuracy: #Squarespace. Squarespace was giving away one free iPhone in the month of July to anyone who had the word Squarespace intheir tweets. As a result we found a lot of tweets like - "I love iPhones #squarespace", "I need an iPhone #squarespace" etc.

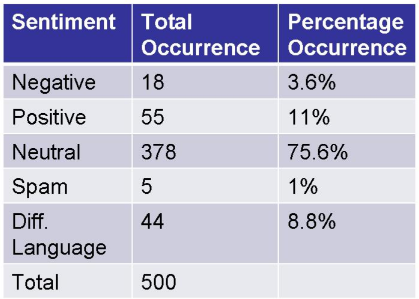
Experiment 2: Machine Learning Techniques

- Classes of Data
- Spam (s)
- Positives (+)
- Negatives(-)
- Neutral(0)



Sentiment	Total Occurrence	Percentage occurrence
Negative	12	0.8%
Positive	150	10%
Neutral	1120	74.67%
Spam	197	13.13%
Diff. Language	21	1.4%
Total	1500	

Figure 6: Training data and the frequency of occurence per class



Sentiment	Total Occurrence	Percentage Occurrence
Negative	18	3.6%
Positive	55	11%
Neutral	378	75.6%
Spam	5	1%
Diff. Language	44	8.8%
Total	500	

Figure 7: Test data and the frequency of occurence per class

- Different language (d)
- Features used
- POS tags
- Emoticons
- Links
- Distance from theme word
- Possession relation
- Love/hate and synonyms
- Wordnet bootstrapped nouns, verbs and adjectives
- Inflections

(Berger, Della Pietra, and Della Pietra 1996) talks about applying Maximum Entropy Modelling to NLP tasks.

Data Statistics

The data we had was quite unbalanced because of the nature of twitter, and the popularity of iPhone. We found very few tweets in our training corpus that had a negative sentiment about iPhone.

Evalutaion Metrics Used

Precision Accuracy of the data retrieved

Recall Correctness of the data retrieved

$$\text{Precision} = \frac{tp}{tp + fp}$$

Figure 8: The formula for finding precision

$$\text{Recall} = \frac{tp}{tp + fn}$$

Figure 9: The formula for finding recall

Machine Learning: Experiment 0

- The dataset was not clearned up at all
- **Training Corpus:** 500 Tweets
- **Test Corpus:** 500 Tweets
- **Features:** 1 gram 5gram
- **Algorithm:** MaxEnt Algorithm, Stanford Parser

Machine Learning: Experiment 1

- The dataset was cleaned
  - Links were replaced by the word LINK
  - Usernames followed by were replaced by PERSON
- **Training data:** 500
- **Test data:** 500
- **Features:** 1 gram 5 gram
- **Algorithm:** MaxEnt Algorithm, Stanford Parser

Machine Learning: Experiment 2

- The dataset was not clearned up at all
- **Training Data:** 1000 Tweets
- **Test Data:** 500 Tweets
- **Features:** 1 gram 4 gram
- **Algorithm:** MaxEnt Algorithm, Stanford Parser

Machine Learning: Experiment 4

- The dataset was not clearned up at all
- **Training Data:** 1500 Tweets
- **Test Data:** 500 Tweets

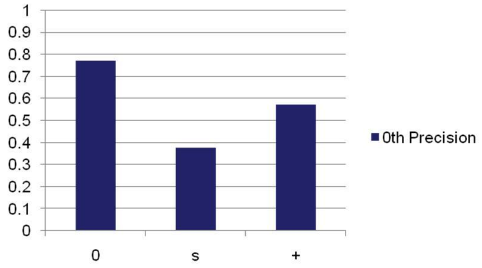


Figure 10: Precision for Experiment 0

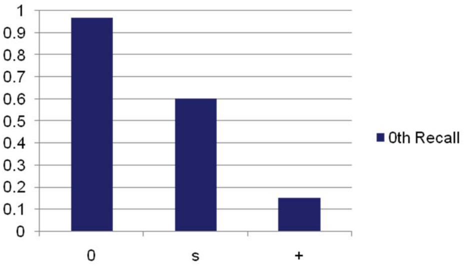


Figure 11: Recall for Experiment 0

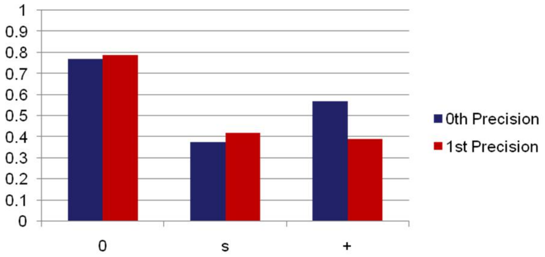


Figure 12: Precision for Experiment 0 and 1

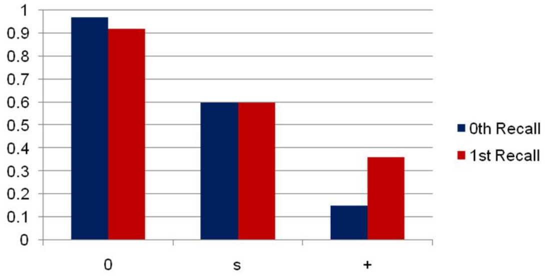


Figure 13: Recall for Experiment 0 and 1

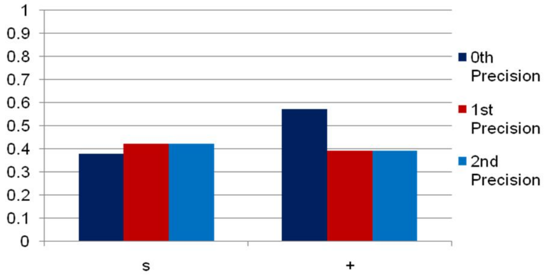


Figure 14: Precision for Experiment 0, 1 and 2

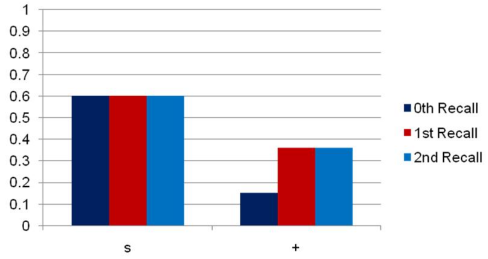


Figure 15: Recall for Experiment 0, 1 and 2

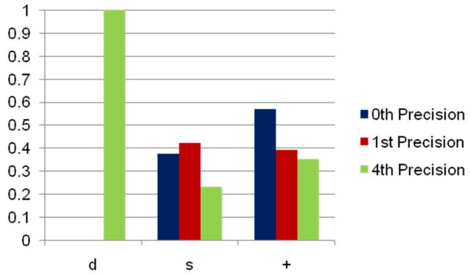


Figure 16: Precision for Experiment 0, 1 and 4

- **Features:** 1 gram 3 gram
- **Algorithm:** MaxEnt Algorithm, Stanford Parser

Machine Learning: Experiment 8a

- only 5 data classes were used instead of the entire tweet
- **Training Data:** 1500 Tweets
- **Test Data:** 500 Tweets
- **Features:**
  - Adjectives
  - Emoticons
  - Links
  - Possession of product want, need, get etc.
- **Algorithm:** Nave bayes, WEKA

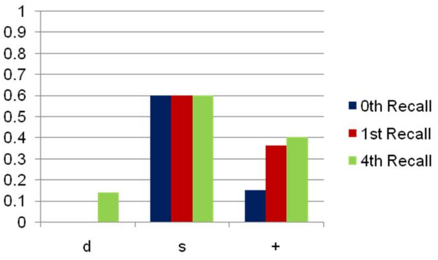


Figure 17: Recall for Experiment 0, 1 and 4

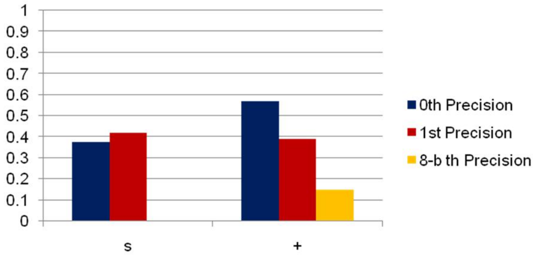


Figure 18: Precision for Experiment 0, 1 and 8a

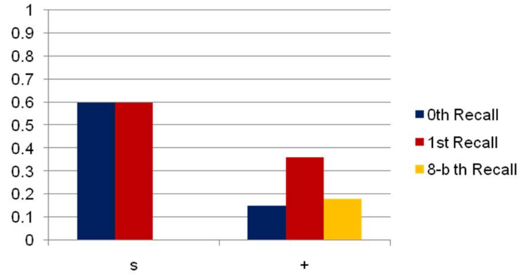


Figure 19: Recall for Experiment 0, 1 and 8a

Machine Learning: Experiment 9

- Replaced some unigrams with their class names
- **Training Data:** 1500 Tweets
- **Test Data:** 500 Tweets
- **Features:**
  - 1 gram 3 gram
  - misspelling correction
  - Pos-neg adjectives
  - Pos-neg emoticons
  - Presence of links
  - Possession
  - love/hate
  - inflection
- **Algorithm:** Maxent model, Stanford Classifier

Future Work

There are many other signals to sentiment, and we would like to use more of these features for classification. Some examples are: using n gram features with Nave Bayes, use POS patterns, use parse information. We would also like to add more training data to our corpus and evaluate the effects of larger corpus on classification. Finally, our classes were extremely unbalanced due to the nature of the domain we selected (iPhone). We would like to select a more balanced domains such as Baseball teams. basketball teams etc. to remove the bias towards dominant classes.

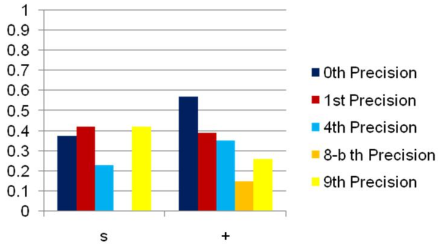


Figure 20: Precision for Experiment 0, 1 and 8a

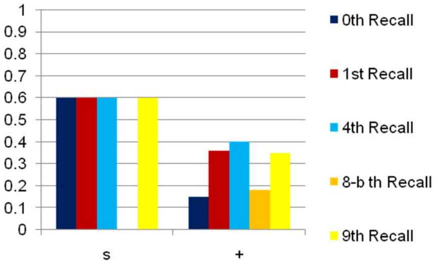


Figure 21: Recall for Experiment 0, 1 and 8a

References

Angela. 2008. Old wine or warm beer target-specific sentiment analysis of adjectives. *Proc.of the Symposium on Affective Language in Human and Machine, AISB 2008 Convention, 1st-2nd April 2008. University of Aberdeen, Aberdeen, Scotland.*

The archivist.

Berger, A. L.; Della Pietra, S. D.; and Della Pietra, V. J. D. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22(1):39–71.

Ding, X.; Liu, B.; and Yu, P. S. 2008. A holistic lexicon-based approach to opinion mining. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, 231–240. New York, NY, USA: ACM.

Du, W., and Tan, S. 2009. An iterative reinforcement approach for fine-grained opinion mining. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 486–493. Boulder, Colorado: Association for Computational Linguistics.

Esuli, A., and Sebastiani, F. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06*, 417–422.

Glance, N. S.; Hurst, M.; Nigam, K.; Siegler, M.; Stockton, R.; and Tomokiyo, T. 2005. Analyzing online discussion for marketing intelligence. In Ellis, A., and Hagino, T., eds., *WWW (Special interest tracks and posters)*, 1172–1173. ACM.

Greene, S., and Resnik, P. 2009. More than words: Syntactic packaging and implicit sentiment. In *Pro-*

ceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 503–511. Boulder, Colorado: Association for Computational Linguistics.

Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *KDD '04: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177. ACM Press.

Huberman, B. A. 2009. Social attention in the age of the web. In *http://www.clir.org/ activities/ digitalscholar2/ huberman11_11.pdf*. Whitepaper.

Jansen, B. J.; Zhang, M.; Sobel, K.; and Chowdury, A. 2009. Micro-blogging as online word of mouth branding. In *CHI EA '09: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, 3859–3864. New York, NY, USA: ACM.

Java, A.; Song, X.; Finin, T.; and Tseng, B. 2007. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, 56–65. New York, NY, USA: ACM.

Kim, S.-M., and Hovy, E. 2006. Automatic identification of pro and con reasons in online reviews. In *Proceedings of the COLING/ACL on Main conference poster sessions*, 483–490. Morristown, NJ, USA: Association for Computational Linguistics.

Krishnamurthy, B.; Gill, P.; and Arlitt, M. 2008. A few chirps about twitter. In *WOSP '08: Proceedings of the first workshop on Online social networks*, 19–24. New York, NY, USA: ACM.

Leskovec, J.; Backstrom, L.; Kumar, R.; and Tomkins, A. 2008. Microscopic evolution of social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 462–470. New York, NY, USA: ACM.

Pang, B.; Lee, L.; and Vaithyanathan, S. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, 79–86. Morristown, NJ, USA: Association for Computational Linguistics.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 347–354.

Zhuang, L.; Jing, F.; Zhu, X.; and Zhang, L. 2006. Movie review mining and summarization. In *Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM)*.

Zhuang, L.; Jing, F.; and Zhu, X.-Y. 2006. Movie review mining and summarization. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, 43–50. New York, NY, USA: ACM.