Sentiment Analysis of Online Spoken Reviews

Verónica Pérez-Rosas, Rada Mihalcea

Computer Science and Engineering
University of North Texas, Denton, TX
veronica.perezrosas@gmail.com,rada@cs.unt.edu

Abstract

This paper describes several experiments in building a sentiment analysis classifier for spoken reviews. We specifically focus on the linguistic component of these reviews, with the goal of understanding the difference in sentiment classification performance when using manual versus automatic transcriptions, as well as the difference between spoken and written reviews. We introduce a novel dataset, consisting of video reviews for two different domains (cellular phones and fiction books), and we show that using only the linguistic component of these reviews we can obtain sentiment classifiers with accuracies in the range of 65-75%.

Index Terms: sentiment analysis, speech transcription, machine learning

1. INTRODUCTION

Online reviews have became an important source of information for both producers and consumers, with companies trying to better understand customer-provided feedback on products and brands, and individual users looking for information to support their everyday purchasing decisions. Given the widespread use of computers and mobile devices, most of which are connected to the Internet, more and more people are sharing their thoughts, feelings, and experiences.

This growing amount of online opinionated information has led to the rapid development of the field of sentiment analysis, which focuses on the automatic identification of opinions, emotions, evaluations, and judgments, along with their polarity (positive or negative). Much of the work to date on sentiment analysis has focused on textual data, such as reviews [1, 2], news articles [3], blogs [4], or Twitter [5]. However, given the accelerated growth of other media on the Web and elsewhere, which includes massive collections of videos (e.g., YouTube, Vimeo, VideoLectures), the ability to address the identification of opinions in the presence of diverse modalities is becoming increasingly important.

In this paper, we address the task of sentiment analysis in online reviews, specifically focusing on the problem of identifying the polarity of spoken opinions. We introduce a novel dataset, consisting of video reviews collected from the ExpoTv website, and we analyze and compare the quality of a sentiment classifier that can be built by using the verbal (linguistic) component of these reviews, obtained through either manual or automatic transcriptions.

Specifically, our goal is to answer the following research questions. First, can we build an automatic sentiment classifier by relying on the linguistic component of these online video reviews, obtained through manual transcriptions? Second, is there a loss in accuracy when the manual transcriptions are replaced with automatic transcriptions? Finally, third, is there a significant difference between the quality of a sentiment analyzer built from spoken reviews as compared to written reviews?

The main contributions of this paper are thus two fold. We introduce and make available a novel dataset consisting of video reviews, which can be used by other researchers to build sentiment classifiers. We also experiment with the verbal component of these reviews, and determine the effect of the quality of the video transcriptions on the accuracy of such sentiment classifiers.

2. RELATED WORK

The techniques developed so far for sentiment analysis have focused primarily on the processing of text, and consist of either rule-based classifiers that make use of opinion lexicons, or data-driven methods that assume the availability of a large dataset annotated for polarity. One of the first lexicons that has been used in polarity analysis is the General Inquirer [6]. Since then, many methods have been developed to automatically identify opinion words [1, 7], as well as n-gram and more linguistically complex phrases [8, 9]. For data-driven methods, one of the most widely used datasets is the MPQA corpus [10], which is a collection of news articles manually annotated for opinions. Other datasets are also available, including polarity datasets covering the domain of movie reviews [2, 11], and a collection of newspaper headlines annotated for polarity [12]. More recently, multi-domain [13] and multi-lingual [14] resources have also been made available.

The increasing number of online multimodal media has made available rich sources of opinionated content. The availability of these resources has motivated the interest in extending the applicability of sentiment analysis tools by incorporating additional data modalities such as speech or video. To date, several exploratory studies addressing this task have been presented using acted data, clean studio recordings, manual transcriptions, and/or expert annotations [15, 16, 17, 18]. Methods like the ones proposed in [19, 20] have approached the music sentiment classification task by combining linguistic and prosodic features extracted from lyrics and audio tracks. These studies have shown the feasibility of combining more than two modalities in the sentiment or emotion recognition task showing that fusing different modalities leads to important improvements (around 8-10% in most cases) over the use of single modalities.

These approaches are however not scalable, because of the human intervention in the process of creating or annotating such data. Instead, the automatic extraction of speech from spoken reviews is preferred in order to analyze the huge volume of data coming from real scenarios, such as home made videos, audio reviews, or logs from call centers. To alleviate the quality issues that are typically associated with these automatic speech recognizers, several directions have been proposed. For instance, Carmelin [21] proposed the evaluation of transcription quality, followed by the selection of the most confident transcription chunks to decrease the amount of noise introduced during the analysis. Metze [22] proposed a word emotional salience method to identify emotion clues in text. Ezzat [23] analyzed the sentiment in automatic transcriptions of agent/customer interactions as a text classification problem using features such as bag of words, term frequency, and keyword extractors, obtaining an accuracy of 66.7% using noisy transcriptions with a word recognition rate of 44%.

Research work has also proposed the addition of acoustic features following the hypothesis that audio clues such as intonations, pauses and prosody can help to alleviate the affect of noise transcriptions. [24] presented an approach to fuse linguistic and acoustic features in order to assess the sentiment in call logs from telephone surveys. [25] presented an evaluation of the role of prosodic clues for sentiment analysis of restaurant spoken reviews.

3. DATASET

To enable our comparative experiments, we compiled a dataset consisting of English video reviews using ExpoTv,[1] which is a public website that provides consumer generated videos. Through this platform users collect unbiased video opinions of products organized in various categories. Our motivation to collect data from this site is the availability of user ratings. For each uploaded video, ExpoTv users provide a star rating for the product they are reviewing (one to five stars). We use this information to assign a sentiment label to each video: videos with four or five stars are labeled as positive, whereas videos with one or two stars are labeled as negative.

To collect the data, we chose two product categories: fiction books and cellphones, which were previously used in sentiment analysis experiments on written text. We then collected the most recent uploaded reviews obtaining 250 videos for fiction books and 150 for cellphones, with an average video length of two minutes.

Transcriptions of the videos in these two collections were obtained using two approaches. First, we collected manual transcriptions by using crowdsourcing via the Amazon Mechanical Turk. Second, we used a speech recognition tool to generate automatic transcriptions.

3.1. Manual Transcriptions

We used the Amazon Mechanical Turk service, which is a crowdsourcing platform provided by Amazon.com. The platform has been heavily used in the past for tasks such as linguistic annotations [26], image labeling [27], translation evaluations [28], and speech transcriptions [29].

A HIT (Human Intelligence Task) was set up on Mechanical Turk, in which workers were provided specific instructions about how to transcribe a video. The guidelines specifically asked for complete, correctly spelled sentences, with punctuation included as needed. The workers were also asked to use filler words, such as "um," "like," "you know." While spam is often an issue with tasks performed by workers on the Mechanical Turk website, we did not receive a significant amount of spam, perhaps due to the fact that this is a widespread task type, and there appears to be a skilled transcriber workforce on Mechanical Turk.

Nonetheless, the transcriptions were manually verified for correctness. We first used simple criteria to accept/reject the transcriptions, such as transcription length (e.g., a transcription that has only one or two lines of text is clearly spam when the corresponding video has a length of 2 minutes). One of the authors then further verified the quality of the transcriptions by checking for the presence of randomly selected utterances from the spoken review inside the transcription. The reviews corresponding to those transcriptions that were rejected were returned to the site for another transcription.

3.2. Automatic Transcriptions

One of the main goals of this paper is to determine the role played by the quality of the transcriptions in the accuracy of a sentiment classifier. Thus, in addition to the manual transcriptions of the reviews, we also experiment with automatic transcriptions, with the aim of making the process of sentiment classification of reviews fully automatic.

There are several speech recognition systems that are commercially or freely available online, such as the Dragon Naturally Speaking tool,[2] or the CMU Sphinx toolkit.[3] However, most of these tools require a training step, and we did not have a training set for our data. We thus opted to use the Google automatic speech recognition engine, which is a ready to use resource available through the Youtube API.[4] We requested automatic transcriptions for our entire dataset, and we obtained captions in the SubRip text format. The API was unable to generate transcriptions for a few of our spoken reviews due to poor quality issues. Thus, after the transcription process, we ended up with a total of 236 and 142 transcription files for the fictions books and the cellphones datasets respectively.

Table 1 shows sample segments of manual and automatic transcriptions. Class distributions and average review length (in number of characters) for the two datasets are shown in Table 2.

Dataset	Instances	Positive	Negative	Review length
Fiction Books	236	131	105	1000
Cellphones	142	78	64	800

Table 2: Class distributions and average review length

3.3. Performance Measures for Automatic Transcriptions

To evaluate the quality of the automatic transcriptions, we use the Sclite tool, which is a freeware resource distributed with the NIST SCTK Scoring Toolkit.[5] Sclite implements an alignment algorithm that evaluates the relation between a hypothesized text (HYP) and a reference (REF) text, and provides statistics such as word recognition rate (WRR), and the number of substitutions, deletions and insertions found while comparing the two sources. Table 3 shows the speech quality statistics for the automatic transcriptions. Since each review is considered as a single sentence, we are not presenting the sentence recognition performance. As it can be observed in the table, the average word error rate of the speech recognition system is 66.4% for both datasets, with similar results for number of substitutions, deletions and insertions; this is expected since most of the videos are recorded in similar settings (i.e., home recordings, surrounding environment noise) with a mix of male and female speak-

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

[2] http://www.nuance.com/dragon

[3] http://cmusphinx.sourceforge.net/

[4] https://developers.google.com/youtube/

[5] http://www.itl.nist.gov/iad/mig//tools/

POSITIVE	NEGATIVE
MANUAL TRANSCRIPTIONS	
Hi. My name's Jane and I'm taking about this book by John. J. Nance's "Final Approach" .... If you like airline thrillers, although I'll put you on the interview seat and make you wonder if you want to fly again.? You will enjoy John J. Nance's "Final Approach"	... I don't have much to say about this. I personally don't like reading that much, but I heard these are really good books. Psych. These books are terrible. Nah I guess, maybe it's just me, but I don't know, but I'm not the kind of guy that exactly likes Harry Potter and likes the magic and likes the, you know, going to school and all that other junk like that.
AUTOMATIC TRANSCRIPTIONS	
My name steven and i'm talking about this book by John Cheney hands final approach ... if you like airline thrillers uh... although put you on the edge of your city and make you wonder at the one fly again you will enjoy John Cheney and says final approach	... don't have much to say about this. I personally don't like being that much book. I heard these are really can books site. These books were terrible night yes mean it's just me. I don't know if I'm not the kind of guy that exactly like terry potter in mike's magic and what's that unit going to school and call that other junk like that.

Table 1: Manual and automatic transcriptions of sample positive and negative spoken reviews

Metric	Cellphones	Fiction Books
Aligned words	46407	64626
% WRR	33.6	33.6
% Substitutions	39.4	40.6
% Deletions	19.1	19.7
%Insertions	5.8	6.0

Table 3: Word recognition performance measures for automatic transcriptions

ers. While the rather low word recognition rate may suggest that the automatic transcriptions would lead to lower sentiment classification performance as compared to the manual transcriptions, through the experiments presented in the next section, we show that combining the automatic speech recognizer output with semantic information obtained from sentiment annotated resources leads to reasonable classification results, with accuracies ranging between 65% and 68%.

4. SENTIMENT ANALYSIS

Our goal in this paper is to perform comparative analyses of sentiment classifiers that can be derived from the linguistic component of spoken reviews. We decided to focus on those features that were successfully used in the past for polarity classification [2, 30]. Specifically, we use: (1) unigram features obtained from a bag-of-words representation, which are the features typically used by corpus-based methods; and (2) lexicon features, indicating the appartenance of a word to a semantic class defined in manually crafted lexicons, which are often used by knowledge-based methods.

Unigrams. We use a bag-of-words representation of the transcriptions to derive unigram counts, which are then used as input features. First, we build a vocabulary consisting of all the words, including stop words, occurring in the transcriptions of the training set. We then remove those words that have a frequency below 10. The remaining words represent the unigram features, which are then associated with a value corresponding to the frequency of the unigram inside each review. Note that we also attempted to use higher order n-grams (bigrams and trigrams), but evaluations on a small development dataset did not show any improvements over the unigram model, and thus all the experiments are run using unigrams.

Semantic Classes. We also derive and use coarse textual features, by using mappings between words and semantic classes. For each semantic class, we infer a feature indicating a raw count of the words belonging to that class. Specifically, we use the following three resources: OpinionFinder (OpF), which is a subjectivity and sentiment lexicon provided with the OpinionFinder distribution [10]; Linguistic Inquiry and Word Count (LIWC), which is a resource developed as a resource for psycholinguistic analysis [31]; and WordNet Affect (WA), which is an affective lexicon created starting with WordNet by annotating synsets with several emotions [32]. Table 4 shows examples of semantic classes from each of these resources.

Class	Words
Opinion Finder	
POSITIVE	abundant, eager, fortunate, modest, nicely
NEGATIVE	abandon, capricious, foul, ravage, scorn
NEUTRAL	absolute, certain, dominant, infectious
LIWC	
OPTIM(ISM)	accept, best, bold, certain, confidence
TENTAT(IVE)	any, anyhow, anytime, bet, betting
SOCIAL	adult, advice, affair, anyone, army, babies
WordNet Affect	
ANGER	wrath, umbrage, offense, temper, irritation
JOY	worship, adoration, sympathy, tenderness
SURPRISE	wonder, awe, amazement, astounding

Table 4: Three word classes from each lexical resource used to derive semantic class features, along with sample words.

5. EXPERIMENTS AND EVALUATIONS

Through our experiments, we address the three main research questions posed in the introduction.

5.1. Can we build an automatic sentiment classifier by relying on the linguistic component of online spoken reviews?

To build the sentiment analysis tool, we use linguistic features consisting of unigrams and semantic classes, as described in Section 4. For the classification, we use the Support Vector Machines (SVM) classifier available in the Weka machine learning toolkit, and run a ten-fold cross validation. Table 5 presents the results obtained using the proposed features sets, as well as combinations among them.

The average classification accuracy obtained with the manual transcriptions ranges between 72-75%. The use of semantic classes appears to help the classification of cellphone reviews, although no improvements are obtained in the case of fiction books. This may be explained by the fact that the book reviews contain more reference to book content (e.g., title, plot, author),

Features	Cellphones		Fiction Books	
	Manual	Automatic	Manual	Automatic
Uni	73.23	62.58	75.42	67.76
Uni+LIWC	74.64	63.94	74.15	67.79
Uni+OpF	72.53	61.90	74.15	66.94
Uni+WA	72.53	62.58	75.00	67.79
Uni+LIWC+OpF+WA	75.35	65.98	72.88	67.37

Table 5: Classification results for manual and automatic transcriptions.

and fewer mentions of actual opinions about the books, which makes the use of the opinion resources less useful.

5.2. Is there a loss in accuracy when the manual transcriptions are replaced with automatic transcriptions?

Our next experiment consists of evaluating the performance of automatically transcribed reviews in the sentiment classification task. We run experiments using the same set of features described above, and once again we use the SVM classifier. The results obtained during these experiments are also presented in Table 5.

When using the automatic transcriptions, we observe a loss in accuracy between 8-10%, which is also explained by the high word error rate measured on these transcriptions. Interestingly, the same pattern is observed in the effect of the semantic class features on the sentiment analysis classifier, where an increase in accuracy is obtained for the cellphones dataset, but no improvements are obtained for the fiction books collection.

5.3. Is there a significant difference between sentiment analysis for spoken and written opinions?

As mentioned before (Section 2), previous work has suggested that text extracted from spoken reviews contains more spontaneous and richer emotional expressions than written reviews and this may provide additional clues for the sentiment analysis task. However, when working with transcriptions, additional challenges appear. For instance, differences in variable utterance lengths and disfluences such as hesitations (e.g. “uh”, “um”), repetitions and corrections [23] introduce additional noise to the analysis, compared with “cleaner” text from written versions.

To explore the differences in sentiment classification when using written or spoken reviews, we decided to empirically compare them using the same machine learning framework. We collected a set of text reviews from Amazon for the same domains (cellphones and books), while preserving the same class distribution and average review length, as shown in Table 2.

Table 6 presents the results obtained using the same linguistic features, for both written and spoken (manually transcribed) reviews. As it can be observed, adding semantic information leads to consistent performance improvement for the written reviews, while for the spoken reviews only the cellphones dataset benefits from these features.

Overall, the results suggest that spoken reviews lead to equal or lower performance as compared to written reviews, which implies that information verbally encoded in multimodal reviews is less informative than the one available in written reviews. One possible explanation for this phenomenon is the fact that when knowing that they are being observed, which is the case of video reviews, people tend to use additional resources such as gestures and intonations, which help them deliver their messages more accurately.

Features	Cellphones		Fiction Books	
	Spoken	Written	Spoken	Written
Uni	73.23	71.12	75.42	84.32
Uni+LIWC	74.64	76.05	74.15	86.01
Uni+OpF	72.53	71.83	74.15	83.89
Uni+WA	72.53	71.83	75.00	84.32
Uni+LIWC+OpF+WA	75.35	75.35	72.88	86.01

Table 6: Classification results for spoken and written reviews.

6. CONCLUSION

In this paper, we addressed the task of sentiment analysis for spoken reviews, with a focus on the verbal component of the reviews. Using a novel dataset, consisting of video reviews from two different domains, we performed evaluations to: (1) determine the accuracy of a sentiment classifier that can be built using only the verbal component of the reviews; (2) measure the role played by the quality of the transcription (manual versus automatic) on the accuracy of the classifier; and (3) compare the performance obtained with written versus spoken reviews. Our findings show that while the use of automatic speech recognition can lead to reasonably accurate sentiment classifiers, with accuracies in the range of 62-68%, the quality of the transcription can nonetheless have a big impact on the sentiment analysis tool, with losses in accuracy of up to 10% for automatic transcriptions as compared to manual transcriptions. Moreover, we found that written and spoken reviews are different in nature, and that the verbal channel of the spoken reviews appears to be less informative than the one in written reviews. The use of semantic classes was found to be consistently useful for written reviews, although their effect on spoken reviews is less clear.

To encourage more research on sentiment classification on video reviews, the datasets introduced are available on request.

ACKNOWLEDGMENTS. This material is based in part upon work supported by National Science Foundation award #0917170 and by DARPA-BAA-12-47 DEFT grant #12475008. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Defense Advanced Research Projects Agency.

7. References

[1] P. Turney, “Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, Philadelphia, 2002, pp. 417–424.

[2] B. Pang and L. Lee, “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts,” in *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, Barcelona, Spain, July 2004.

[3] K. Balog, G. Mishne, and M. de Rijke, “Why are they excited? identifying and explaining spikes in blog mood levels,” in *Proceedings of the 11th Meeting of the European Chapter of the As sociation for Computational Linguistics (EACL-2006)*, 2006.

[4] N. Godbole, M. Srinivasaiah, and S. Sekine, “Large-scale sentiment analysis for news and blogs,” in *International Conference on Weblogs and Social Media*, Denver, CO, 2007.

[5] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, “Target-dependent Twitter sentiment classification,” in *Proceed-*

ings of the Association for Computational Linguistics (ACL 2011), Portland, OR, 2011.

[6] P. Stone, *General Inquirer: Computer Approach to Content Analysis*. MIT Press, 1968.

[7] M. Taboada, J. Brooke, M. Tofiloski, K. Voli, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, no. 3, 2011.

[8] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in *Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, 2003, pp. 105–112.

[9] H. Takamura, T. Inui, and M. Okumura, "Latent variable models for semantic orientations of phrases," in *Proceedings of the European Chapter of the Association for Computational Linguistics*, 2006.

[10] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 39, no. 2-3, pp. 165–210, 2005.

[11] A. Maas, R. Daly, P. Pham, D. Huang, A. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the Association for Computational Linguistics (ACL 2011)*, Portland, OR, 2011.

[12] C. Strapparava and R. Mihalcea, "Semeval-2007 task 14: Affective text," in *Proceedings of the 4th International Workshop on the Semantic Evaluations (SemEval 2007)*, Prague, Czech Republic, 2007.

[13] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *Association for Computational Linguistics*, 2007.

[14] C. Banea, R. Mihalcea, and J. Wiebe, "Multilingual subjectivity: Are more languages better?" in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China, August 2010, pp. 28–36. [Online]. Available: http://www.aclweb.org/anthology/C10-1004

[15] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572 – 587, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320310004619

[16] S. Raaijmakers, K. Truong, and T. Wilson, "Multimodal subjectivity analysis of multiparty conversation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, 2008, pp. 466–474.

[17] Z. Zhihong, M. P. G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *PAMI*, vol. 31, no. 1, 2009.

[18] L. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the International Conference on Multimodal Computing*, Alicante, Spain, 2011.

[19] J. Zhonga, Y. Chenga, S. Yanga, and L. Wena, "Music sentiment classification integrating audio with lyrics," 2012.

[20] R. Mihalcea and C. Strapparava, "Lyrics, music, and emotions," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 590–599. [Online]. Available: http://www.aclweb.org/anthology/D12-1054

[21] N. Camelin, F. Bechet, G. Damnati, and R. De Mori, "Detection and interpretation of opinion expressions in spoken surveys," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 2, pp. 369 –381, feb. 2010.

[22] F. Metze, A. Batliner, F. Eyben, T. Polzehl, B. Schuller, and S. Steidl, "Emotion recognition using imperfect speech recognition." in *INTERSPEECH*, T. Kobayashi, K. Hirose, and S. Nakamura, Eds. ISCA, 2010, pp. 478–481. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2010.html#MetzeBEPSS10

[23] S. Ezzat, N. Gayar, and M. Ghanem, "Investigating analysis of speech content through text classification," in *Soft Computing and Pattern Recognition (SoCPaR), 2010 International Conference of*, dec. 2010, pp. 105 –110.

[24] F. Metze, T. Polzehl, and M. Wagner, "Fusion of acoustic and linguistic features for emotion detection," in *Semantic Computing, 2009. ICSC '09. IEEE International Conference on*, sept. 2009, pp. 153 –160.

[25] F. Mairesse, J. Polifroni, and G. Di Fabbrizio, "Can prosody inform sentiment analysis? experiments on short spoken reviews," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, march 2012, pp. 5093 –5096.

[26] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng, "Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, 2008.

[27] C. Leong, R. Mihalcea, and S. Hassan, "Text mining for automatic image tagging," in *International Conference on Computational Linguistics*, Beijing, China, August 2010, pp. 647–655. [Online]. Available: http://www.aclweb.org/anthology/C10-2074

[28] M. Bloodgood and C. Callison-Burch, "Using Mechanical Turk to build machine translation evaluation sets," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[29] M. Marge, S. Banerjee, and A. I. Rudnicky, "Using the Amazon Mechanical Turk for transcription of spoken language," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010.

[30] J. Wiebe and E. Riloff, "Creating subjective and objective sentence classifiers from unannotated texts," in *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005) (invited paper)*, Mexico City, Mexico, 2005.

[31] J. Pennebaker and M. Francis, "Linguistic inquiry and word count: LIWC," 1999, erlbaum Publishers.

[32] C. Strapparava and A. Valitutti, "Wordnet-affect: an affective extension of wordnet," in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, 2004.