AUTOMATIC SENTIMENT EXTRACTION FROM YOUTUBE VIDEOS

Lakshmish Kaushik, Abhijeet Sangwan, John H.L. Hansen

Center for Robust Speech Systems (CRSS), Eric Jonsson School of Engineering,
The University of Texas at Dallas (UTD), Richardson, Texas, U.S.A.
{lakshmish.kaushik, abhijeet.sangwan, john.hansen}@utdallas.edu

ABSTRACT

Extracting speaker sentiment from natural audio streams such as YouTube is challenging. A number of factors contribute to the task difficulty, namely, Automatic Speech Recognition (ASR) of spontaneous speech, unknown background environments, variable source and channel characteristics, accents, diverse topics, *etc*. In this study, we build upon our previous work [5], where we had proposed a system for detecting sentiment in YouTube videos. Particularly, we propose several enhancements including (i) better text-based sentiment model due to training on larger and more diverse dataset, (ii) an iterative scheme to reduce sentiment model complexity with minimal impact on performance accuracy, (iii) better speech recognition due to superior acoustic modeling and focused (domain dependent) vocabulary/language models, and (iv) a larger evaluation dataset. Collectively, our enhancements provide an absolute 10% improvement over our previous system in terms of sentiment detection accuracy. Additionally, we also present analysis that helps understand the impact of WER (word error rate) on sentiment detection accuracy. Finally, we investigate the relative importance of different Parts-of-Speech (POS) tag features towards sentiment detection. Our analysis reveals the practicality of this technology and also provides several potential directions for future work.

Index Terms: Audio sentiment detection, Reviews, Maximum Entropy, POS tagging, KALDI, NLP, ASR, YouTube

1. INTRODUCTION

Social networking applications such as Twitter, Facebook, YouTube, etc. are popularly used to express one's sentiment and/or opinion on a variety of topics. A large number of these applications rely on text as the main medium of communication. However, websites such as YouTube use video/audio as the primary source of communicating information. For example, "unboxing" is a very popular theme on YouTube where users express their opinion and sentiment about products while unpacking[1] and experiencing the product for the first time. Sentiment systems that can crawl and mine these information resources can assist in establishing the popular sentiment or the "word of mouth" on a large range of topics. Such information can be tremendously useful to businesses and consumers alike. Text-based sentiment analysis has been well researched and numerous techniques that mine reviews for opinions have been developed [1-4]. However, audio-based sentiment analysis remains under explored. Recently, we had shown that audio sentiment extraction with good accuracy is possible using a combination of NLP (natural language processing) and ASR (automatic speech recognition) techniques [5]. Particularly, we had demonstrated the capability of automatically predicting the polarity of sentiment (positive or negative). First, audio was extracted from the YouTube video and then converted to text using the ASR system, and finally the text-based sentiment system predicted the sentiment polarity. The text-based sentiment system used parts-of-speech tagging technique to automatically extract text-features, which were then employed in a maximum entropy based classification system to predict sentiment polarity.

In this study, we build on our previous work. Firstly, we adopt a more powerful speech recognition system in order to boost the system accuracy. Our previous language model was trained on a combination of conversational telephony and web based data. In this study, we have incorporated a focused sentiment vocabulary in our dictionary, and our language model includes text from reviews. These modifications improve our speech recognizer, as the system is now more adept at capturing textual sentiment features (thereby improving the overall system accuracy).

Our text based sentiment estimation system uses Maximum Entropy (ME) classifier and POS tagged text features. In this study, we have also increased the amount and diversity of training data used to build the ME sentiment engine with an intention of developing a more accurate system. Furthermore, we have also observed that the proposed ME model contains a large number of features (as many as 1.4 million). We have seen that a number of these features tend to be ambiguous as they show weak affinity towards either sentiment. Hence, we develop a 0pruning technique that removes ambiguous features based on the intuition that removing such weak features should reduce model complexity with minimal impact on accuracy. Additionally, removing a large number of ambiguous features also helps in reducing the size of the ASR vocabulary (which can have a positive impact on ASR performance).

Additionally, we have also increased the size of our YouTube evaluation set. The new system (which is the result of the mentioned improvements) gives an overall improvement of 10% absolute in sentiment detection accuracy over our previous system. Finally, we also present analysis that sheds more light on the interplay between the ASR and NLP components of the sentiment detection system. Particularly, we conduct experiments to (i) understand the role of WER on sentiment

*This material is based upon work supported in part by AFRL under contract FA8750-12-1-0188, by NSF under Grant 1218159, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen. Opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of0 the National Science Foundation, or AFRL.

[1] http://bit.ly/YLha0h

detection accuracy and (ii) the relative importance of different categories of POS tagged textual features towards system performance.

2. DATABASE COLLECTION

2.1 Sentiment text database.

In order to develop a more comprehensive system, we have significantly increased the size and diversity of our training dataset. The following data sources have been used in this study:

a) Amazon Product Reviews [7],
b) Pros & Cons database [8],
c) Comparative Sentence Set Database [9],
d) Opinion Lexicon Database [7],
e) Scale database [10], and
f) R-T polarity database [7].

The total dataset contains about 6 million review comments on a wide variety of products such as books, household wares, electronic goods, apparels, movies etc. Some of these datasets (a, b, and c) are publicly available [7]. Additionally, we also used an Amazon review scraper tool to download a large number of reviews [11]. A more detailed explanation of these datasets can be found in [7,8,9,10].

2.2 YouTube audio database.

YouTube videos [12] are an ideal choice for system evaluation since speakers follow a natural and spontaneous speaking style while sharing their opinion on a wide variety of topics. We have collected a set of 85 videos (44 positive and 41 negative) that contain people speaking spontaneously. These videos cover a wide range of topics including product reviews, movies, social issues and political opinions. Our YouTube sentiment database contains 55 male speakers, 30 female speakers, and 5 videos with multiple speakers. The average duration of the videos is about 5 minutes with individual video durations ranging from 2 to 15 minutes. The total duration of the evaluation dataset is about 7.5 hours. The audio quality, recording equipment, channel characteristics, and accents/dialects vary across videos. These videos can be accessed via a YouTube playlist: http://bit.ly/YAgoYU.

Three listeners listened to the above videos to establish the sentiment ground truth. The listeners were asked to judge the videos for positive or negative sentiment. Later, their judgment was averaged to yield the final decision. This decision was considered as ground truth for the system evaluation experiments.

3. SENTIMENT MODEL GENERATION

Figure 1 gives the methodology for developing Text based Sentiment models. First, the raw text from the database mentioned in Section 2.1 are processed for parts of speech (POS) tagging. Subsequently, these features are used to train sentiment models using the Maximum Entropy (ME) method. But the models obtained are not optimal. It contains a huge set of redundant features which might not depict sentiment decisively. Hence the feature set can be reduced to obtain a compact and more efficient sentiment models. To achieve that we perform iterative feature reduction during training. Finally, the sentiment models are used in conjunction with ASR to perform sentiment detection on YouTube videos. Following sections explain in details the implementation of each of the blocks in Figure 1.

3.1. Textual Feature Extraction using Parts of Speech (POS) Tagging

We use parts of speech (POS) tagging to automatically determine textual features for sentiment detection. In what follows, we explain this process. Table 1 shows typical sample reviews, where user rating and textual comments are captured. Here, the rating is treated as the sentiment ground truth for the corresponding textual comment.

In general, words and/or word combinations formed by adjective (JJ), Verb (V*), Adverb (RB*) and Noun (N) tend to capture the expressed sentiment. For example, readers can review the words in bold face in Table 1. Words and word combinations such as "excellent", "great speed" express positive sentiment, and "horrible", "nonsensical rambling" suggest negative sentiment. Using POS tagging, the mentioned words/word combinations can be extracted by parsing the input text. In this study, we have used the Stanford's Log-linear POS Tagger for POS tagging [13].

Table 1: *Example review text with corresponding 5-point rating.*

Rating	Review
★★★★★	The Phones Works **Excellent**, **great speed**, the cam get really **great quality**. Android+HTC sense is a **nice interface**.
★☆☆☆☆	**Horrible** piece of **Garbage**. This is one of the **worst pieces** of writing I have ever come across. I **thoroughly discourage** anyone even picking up this book, and I would rather sit through a screening of Battlefied Earth than subject myself this **nonsensical rambling**.

In general, features derived from noun combinations are more likely to be product features, whereas adjective based features more likely to capture the sentiment directed towards the product's features. Similarly, verbs and adverbs based features are likely to capture the product functionality and opinions. We identify the following POS tag combinations namely, JJ, JJ-JJ, JJ-V*, V-JJ, RB-JJ, RB-V, V*-RB, JJ-NN, NN-JJ, JJ-JJ-NN, V*-NN, JJ-IN-NN and extracted these as features. For example, Table 2 shows the corresponding extracted textual features for the review comments in Table 1.

Using the mentioned technique, textual features are extracted for all comments in the training dataset. Following this process, we obtain a parallel corpus of sentiment polarity (from ratings) and textual sentiment features (from review). A threshold of 2.5 is used to convert the 5-point rating scale into positive and negative sentiment, *i.e.*, a rating of greater than or equal to 2.5 is considered positive, and a rating of lesser than 2.5 is considered negative. For example, Table 2 shows the binary sentiment derived from the ratings in Table 1.

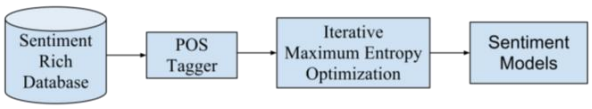

Figure 1: *Text based Sentiment model development using Maximum entropy method.*

From the datasets mentioned in Sec. 2.1, we have generated features for 6 million reviews. Out of this 4 million, 1 million and 1 million reviews are used for training, development and evaluation, respectively.

Table 2: *Corresponding sentiment features and polarity extracted from the example reviews and ratings, respectively (shown in Table 1). The sentiment features and polarity are used for training sentiment models.*

Sentiment Polarity	Sentiment Feature
1	excellent great_speed great_quality nice_interface.
0	Horrible Garbage worst_pieces thoroughly_discourage nonsensical_rambling.

We also develop several variations of the sentiment engine, where different feature sets are used to train the ME classifier. These variations are as follows:

1. All POS tags combinations are employed as features,
2. All POS tags combinations are employed except features that use noun POS tags,
3. All POS tags combinations are employed except features that use noun and/or verb POS tags, and
4. All POS tags combinations are employed except features that use verb POS tags.

The mentioned variations in the sentiment detection system allows us to compare the effectiveness of the different feature sets.

3.2 Maximum Entropy Modeling

The proposed system uses Maximum Entropy (ME) modeling technique [14-16] to determine comment sentiment polarity given textual sentiment features as input. Let y_j be the j^{th} sentiment where $y_j \in Y$ and $Y \equiv \{positive, negative\}$ is the set of sentiment polarities. Let x_k be k^{th} textual sentiment feature, then function f_i is defined as:

$$f_i(x_k, y_j) = \begin{cases} 1 & if\ x_k\ is\ present\ in\ text\ comment \\ 0 & otherwise \end{cases}$$

The functional definition above hypothesizes a relation between a feature present in the review text, and the corresponding review ratings. Applying evidence based modeling technique like ME the relationship can be estimated quantitatively. The ME technique can predict the rating of the review y_j from features x_k by using:

$$p(y_j|x_k) = \frac{1}{Z_\lambda(x)} \sum_{i=1}^{N} \exp\left(\lambda_{ij} f_i(x_k, y_j)\right) \quad (1)$$

where, $Z_\lambda(x)$ is a normalizing term, and λ_{ij} are weights assigned to the f_i. The training data described in Sec 2.1 is used to develop the ME model for this study.

4. ITERATIVE SENTIMENT FEATURE REDUCTION

We have observed that the ME based sentiment model (developed using the technique outlined in Sec. 3.2) tends to generate a large number of features. Furthermore, the number of unique words in this textual feature set tends to be very high (around 250K). This has implication on ASR design where we must now run decoding with a very large dictionary and language model (LM). This has the potential to reduce accuracy and produce a slower system. Hence, it is worth investigating if the complexity of the sentiment model can be reduced in order to gain efficiency and effectiveness.

We adopt a simple yet effective technique to reduce sentiment model complexity. From the baseline model, the probabilities of positive and negative sentiment for every textual sentiment feature can be estimated independently. In this manner, the effectiveness of each feature can be studied in isolation. It is reasonable to assume that some features would predict positive or negative sentiment more strongly than others. For example, lets assume that "A" and "B" are sentiment features and probability of positive sentiment given A is unity, and probability of positive sentiment given B is 0.5. Now, it can be inferred that B is an ambiguous feature since it fails to decisively select positive or negative sentiment, and A is unambiguous since it clearly predicts positive sentiment. In the proposed feature pruning scheme, it is our intention to remove as many ambiguous feature as possible and retain unambiguous features, thereby maintaining classification accuracy and reducing model complexity.

Now, we formalize our method. We use the following criterion to determine if a feature is ambiguous or unambiguous:

$$x_k = \begin{cases} unambigious & if\ max\left(p(y|x_k)\right) \geq 0.55 \\ ambigious & if\ max\left(p(y|x_k)\right) < 0.55 \end{cases}$$

Using this criterion, ambiguous features are removed and the sentiment model is retrained using unambiguous features alone. Once the new model is trained, then ambiguous features are again identified, removed and a new model is trained. This process is repeated and the performance of the model generated at every iteration is measured using a held out evaluation set. In this manner, we can track feature set cardinality (*i.e.*, total number of features) with performance accuracy for every iteration.

It is useful to note that several stopping criteria can be adopted for the iterative procedure mentioned here. In our study, the process is stopped when the difference in the feature set cardinality between successive iterations is no more than 500 features. Using this technique, we were able to reduce the number of textual features from 1.4 million features in our baseline model to 280K in the final model.

It is also worth mentioning that additional criterion (in conjunction with ambiguity of textual feature) can be applied towards obtaining an effective sentiment model. For example, "detectability" of textual feature by ASR and/or frequency of occurrence of the textual feature for a given domain.

5. AUTOMATIC SPEECH RECOGNITION SYSTEM

We describe acoustic and language model development for the proposed sentiment detection system.

5.1 Acoustic and Language Models

We used standard triphone based Hidden Markov Models (HMMs) for this study. We also used standard MFCC (Mel Frequency Cepstral Coefficients) features with delta and delta-delta coefficients. In order to compute the MFCCs, 24 mel filter banks spanning frequencies from 25Hz to 3800Hz were employed. Utterance level cepstral mean normalization (CMN) and speaker level cepstral variance normalization (CVN) was also applied. Finally, LDA/MLLT (linear discriminant analysis/maximum likelihood linear transform) was applied to the cepstral features. Speaker adaptive training (SAT) using fMLLR (feature space MLLR) was used to obtain the final acoustic models. In this study, the acoustic models were trained on a mixture of switchboard and fisher corpora (totaling up to

600 hours of training data). Also, we used the Kaldi recognition toolkit to build the acoustic models for this study [6].

A trigram language model was trained using the following data sources: (i) Switchboard, (ii) Fisher (iii) UW191 [17] (191M words collected from the web by the University of Washington), and (iv) sentiment datasets mentioned in Sec. 2.1. The recognition lexicon contained 90K words.

During decoding, we executed two rounds of fMLLR transform estimation, before using the second pass fMLLR transform for rescoring the decoded lattices. Finally, the 1-best hypothesis was captured and passed onto the text based sentiment system.

5.2 ASR based sentiment detection in YouTube videos

Figure 5 shows the overall system used for detecting sentiment from YouTube videos. As a first step, audio is stripped from the YouTube videos. The recognition setup described in Sec. 5.1 is used to obtain 1-best speech transcripts from the stripped audio data. After decoding, the text is parsed by the POS tagger to obtain textual sentiment features. Finally, ME based sentiment models are used to detect sentiment polarity given the sentiment features.

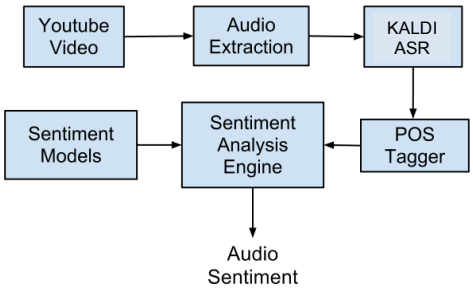

Figure 2: *Proposed sentiment detection system for audio streams using ASR to convert speech to text and text based sentiment detection system to extract sentiment.*

6. RESULTS AND DISCUSSION

Here, we present the analysis of the text and audio based sentiment detection system. Section 7.1 explains analysis and results obtained from the Text based Sentiment detection and section 7.2 presents the sentiment detection results for YouTube videos.

6.1 Text based sentiment detection system

6.1.1 Accuracy of the proposed text based sentiment detection system

In order to benchmark the proposed ME based sentiment detection system, we compare our approach to two standard techniques: (i) Naïve Bayes proposed in [9] and (ii) Support Vector Machine (SVM) approach proposed in [18]. In [9], the authors used the comparative sentences dataset for evaluation. Similarly, we also use the comparative sentences dataset for evaluation of the proposed ME system and report the performance in Table 3. It is noted that 'Reviews', 'Articles' and 'Forums' are three sub-tasks in the comparative sentences dataset evaluation. From Table 3, it can be observed that the proposed ME system consistently outperforms the Naïve Bayes approach by an absolute margin of 8-to-9%.

Additionally, Table 4 shows the comparison of our technique to a more recent technique that uses Passive Aggressive (Online SVM) technique (designed for substantially large training and testing sets). In [18], the authors use a dataset that contains around 180K training samples and 125K evaluation samples. This dataset mainly consists of product reviews from CNet, Amazon and Yahoo. Table 4 shows the performance of the proposed ME system on this dataset. As seen from the table, the performance of the proposed ME technique is slightly inferior to the SVM technique. While we have chosen to adopt the ME technique for the remainder of this study, this result clearly demonstrates that we can further improve our text-based sentiment detection system.

Table 3: *Text based sentiment detection accuracy from ME sentiment models compared with a standard Bayesian technique.*

Data Sets	Precision		Recall		F-Score	
	Naive Bays	Entropy	Naive Bays	Entropy	Naive Bays	Entropy
Reviews	0.84	0.92	0.8	0.90	82%	91%
Articles	0.75	0.88	0.8	0.84	77%	86%
Forums	0.73	0.86	0.83	0.86	78%	86%

Table 4: *Text based sentiment detection accuracy from ME sentiment models compared with Online SVM technique.*

Method	**Precision**	**Recall**	**F-Score**
Passive Aggressive Technique (Online SVM)	0.9022	0.8991	90.07%
Iterative Maximum Entropy	0.8732	0.8498	86.15%

6.1.2 Impact of Iterative Feature Reduction on ME models

For this experiment, the iterative reduction scheme was applied to all four sentiment model variations, namely, (i) With Noun, No Verb, (ii) With Noun, (iii) Without Noun, and (iv) Without Noun, No Verb. The results are shown in Fig. 3. The training and evaluation for this experiment was executed within the text domain.

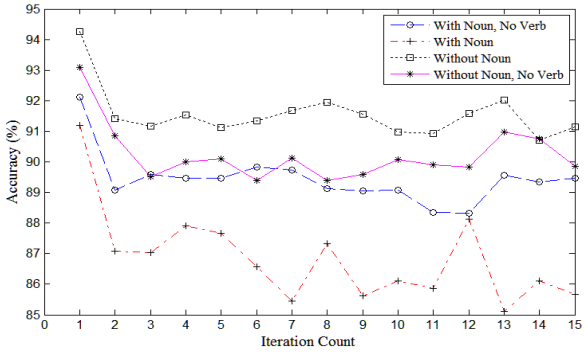

Figure 3: *Sentiment detection performance with iterative feature reduction technique during training the sentiment models using the proposed ME technique.*

From Fig. 3, it is observed that the Without Noun system delivers the best performance. On the other hand, the With Noun system performs the worst. We suspect that having more noun features makes the system more domain dependent. This hurts the generality of the sentiment model which in turn results in poorer performance.

With increasing iterations, the performance of all systems is observed to first drop and then stabilize. The initial drop is different for different systems (between 3-to-5%), but nominal compared to the drop in number of features (which is 1.4 million to 280K for No Noun system).

6.1.3 Impact of ASR Word-Error-Rate (WER) on Sentiment Detection

In order to understand the impact of ASR WER (word error rate) on the proposed audio sentiment detection system, we construct an experiment where controlled amounts of substitution errors are introduced within the clean text to generate various noisy versions. In this manner, we can control the exact amount of WER, and the corresponding sentiment detection accuracy is measured.

One more parameter to control carefully while performing this experiment is the review length in words. For example, at 50% WER, a review comment of 100 words would still have 50 correct words, but a review comment of 10 words would only have 5 correct words. Therefore, it is likely that the information loss at 50% WER is far more damaging for shorter reviews than longer ones. To understand this phenomenon better, we form three different evaluation sets of similar review lengths. Particularly, our smallest review category consists of reviews that are 5-to-50 words long. The next category contains reviews that are 50-80 words long, and the final category contains reviews that are 80-120 words long. Each category contains 100K reviews, with 50K positive and negative comments. We simulate WERs in each category and measure the corresponding sentiment detection accuracy separately.

Fig. 4 shows the variation in sentiment detection accuracy with increasing WER for the 3 mentioned review categories. As expected, it is seen that the sentiment detection accuracy falls with decreasing WER. Also, the longest review comments are most robust to WER. The greatest jump in accuracy for all three categories is observed when the WER falls from 100% to 80%. Subsequent falls in WER (in increments of 20%) yield declining increase in sentiment detection accuracy. The key take away from this experiment is that sentiment detection is quite robust to WER, especially for longer comments. Conversely, the challenge in automatic sentiment detection in audio is perhaps in shorter comments.

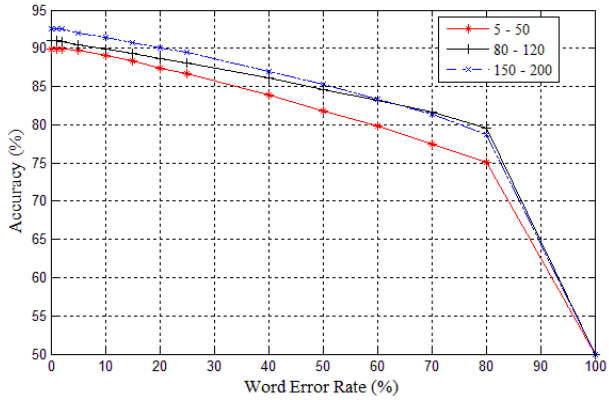

Figure 4: *Simulated WER graph for text based system where correct words are randomly selected and substituted (thereby introducing substitution errors). Four different comment lengths are considered for evaluation. It can be seen that longer comments are more robust to ASR errors.*

6.2 Sentiment detection for YouTube videos

Table 4: S*entiment detection accuracy of YouTube videos for different POS word-combination.*

POS Combination Scenario	Sentiment Accuracy
Without Noun	88%
Without Noun No verb	84%
With Noun	83%
With Noun No Verb	81%

6.2.1 Sentiment Accuracy

Table 4 shows the sentiment detection accuracy for the proposed system, for all 4 variations of the ME system (mentioned in Sec. 3.1). It can be observed that Without Noun system gives the highest sentiment detection accuracy (88% accuracy). It is interesting to note that the best performance with our older system [5] was 78% on the current YouTube evaluation dataset. This constitutes an absolute improvement of 10% in accuracy. Additionally, it is also seen that the With Noun No Verb systems gives the poorest performance (81% accuracy). The results seem to suggest that Verb information is more useful towards sentiment classification than Noun information. It is also possible that Noun information tends to be domain dependent, which may reduce its effectiveness in a general evaluation.

6.2.2 DET Curve analysis

Figure 5 presents the results for sentiment detection for the YouTube evaluation dataset in form of DET (detection error tradeoff) curves. In order to construct the DET curve, we use soft decisions from the ME model (as opposed to using hard decisions), namely, the probabilities of positive and negative sentiments, and a variable threshold. Additionally, we also show the DET curves for all 4 variations of ME system.

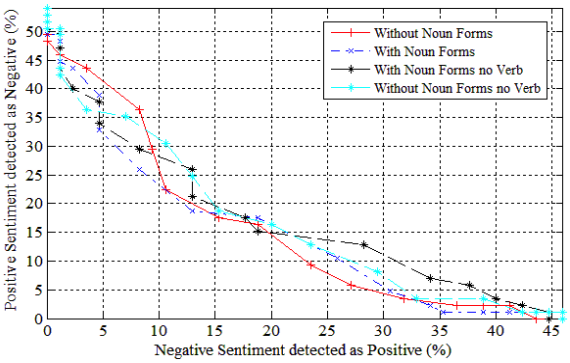

Figure 5: *DET curves for 4 variations of the ME system. Without Noun system gives the best EER (equal error rate) result.*

From Figure 5, it can be seen that the Without Noun system gives the best EER performance, which is approximately 17%. It is useful to note that the EER is 5% more than the smallest error rate obtained for the Without Noun system (which is 12% corresponding to 88% accuracy). Also, it can be seen that the Without Noun and With Noun systems seem to be relatively better at detecting positive and negative sentiments, respectively. Interestingly the EER performance of all 4 systems is nearly the same. However, the Without Noun system has the least number of textual features when compared to the other systems.

7. CONCLUSIONS

In this study, we have presented several improvements over our old audio-based sentiment detection system [5]. Firstly, we have used significantly more data from diverse sources for training our text-based ME sentiment models. Next, we have also proposed a new method that iteratively prunes ambiguous features from the ME based sentiment model. This method delivers the benefits of building more efficient ME sentiment models, a smaller vocabulary, and more focused ASR vocabulary/language model. Additionally, this method allows us to continue to increase our training dataset while managing model complexity. Next, we have used a more powerful KALDI based speech recognition engine that uses SAT (speaker adaptive training) acoustic models with a bigger (and more application focused) language model. The combination of these improvements have delivered an absolute improvement of 10% in sentiment detection on a difficult YouTube evaluation dataset. Finally, we have also presented new analysis that helps in understanding the impact of WER on sentiment detection and the relative contribution of different POS tags based textual features towards accuracy.

Audio based sentiment detection remains an unexplored area of application within speech and language technology. Hence, several avenues remain open for future work. While this study has focused on determining the unknown sentiment in a YouTube video, it would also be interesting to automatically extract the object at which the sentiment is directed. For example, in the comment "the food was tasteless", the negative sentiment "tasteless" is directed at the object "food". Additionally, it would also be interesting to automatically extract demographic information such as age, gender *etc.* about the speaker of the sentiment. Furthermore, it should also be possible to use speech processing techniques to automatically extract emotion/stress related information from the signal. Finally, more work is also required to understand the interaction between speech processing, ASR, and NLP components and the impact they have on the overall design of the system.

8. REFERENCES

[1] B. Liu. "Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing", Second Edition, 2010.

[2] B Pang and Lee "Foundations and Trends in Information Retrieval" 2(1-2), pp. 1–135, 2008.

[3] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 168–177, 2004.

[4] N. Jindal, and B. Liu. "Opinion Spam and Analysis." Proceedings of the ACM Conference on Web Search and Data Mining (WSDM), 2008.

[5] L. Kaushik, A. Sangwan, and J.H.L. Hansen, "Sentiment Extraction from Natural Audio Streams," ICASSP'13, Vancouver, Canada, 2013.

[6] D. Povey et. al., "The Kaldi speech recognition toolkit," ASRU'11, Hawaii, U.S.A., 2011.

[7] www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[8] M. Ganapathibhotla, B. Liu: "Mining Opinions in Comparative Sentences." COLING, pages 241-248, 2008

[9] N. Jindal, B. Liu: "Identifying comparative sentences in text documents." SIGIR, pages 244-251, 2006.

[10] B. Pang and L. Lee. 2005. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In ACL, pages 115-124, 2005

[11] Amazon reviews downloader and parser, http://esuli.it

[12] www.youtube.com

[13] K. Toutanova, D. Klein, C. Manning, and Y. Singer, "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network". In Proc. of HLT-NAAC, pp. 252-259, 2003.

[14] A. Ratnaparkhi, "A Maximum Entropy Model for Part-of-Speech Tagging", In Proc. of the Empirical Methods in Natural Language Processing, pp. 133-142, 1996

[15] C. Maxent, "mathematics, and information theory, Maximum Entropy and Bayesian Methods." Kluwer Academic Publishers, 1996.

[16] R. Rosenfeld, "Adaptive Statistical Language Modeling: A Maximum Entropy Approach." PhD thesis, Carnegie Mellon University, 1994.

[17] http://ssli.ee.washington.edu/ssli/projects/ears/WebData/web_data_collection.html

[18] C. Hang, V. Mittal, and M. Datar. "Comparative experiments on sentiment classification for online product reviews." In AAAI, vol. 6, pp. 1265-1270. 2009