SENTIMENT EXTRACTION FROM NATURAL AUDIO STREAMS

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

Automatic sentiment extraction for natural audio streams containing spontaneous speech is a challenging area of research that has received little attention. In this study, we propose a system for automatic sentiment detection in natural audio streams such as those found in YouTube. The proposed technique uses POS (part of speech) tagging and Maximum Entropy modeling (ME) to develop a text-based sentiment detection model. Additionally, we propose a tuning technique which dramatically reduces the number of model parameters in ME while retaining classification capability. Finally, using decoded ASR (automatic speech recognition) transcripts and the ME sentiment model, the proposed system is able to estimate the sentiment in the YouTube video. In our experimental evaluation, we obtain encouraging classification accuracy given the challenging nature of the data. Our results show that it is possible to perform sentiment analysis on natural spontaneous speech data despite poor WER (word error rates).

Index Terms-Sentiment detection, Speech Recognition, Maximum entropy, YouTube, Amazon product review data.

1. INTRODUCTION

In the contemporary world, internet based multimedia has become the main source of presenting one's opinion. This is primarily because regular Internet users have a wider sphere of influence through larger social circles. It is no surprise that among Internet users peer recommendation forms one of the most important sentiment or opinion [1].

"YouTube" is one such enormous social circle where people visit regularly for gathering information or opinions about various topics. In a large proportion of these videos, people depict their opinions about products, movies, social issues, political issues, etc. The capability of detecting the sentiment of the speaker in the video can serve two basic functions: (i) it can enhance the retrieval of the particular video in question, thereby, increasing its utility, and (ii) the combined sentiment of a large number of videos on a similar topic can help in establishing the general sentiment. It is important to note that automatic sentiment detection using text is a mature area of research, and significant attention has been given to product reviews [2-8]. In this study, we focus our attention on automatic sentiment detection in YouTube videos based on audio analysis. We focus on YouTube because the nature of speech in these videos is more natural and spontaneous which makes automatic sentiment processing challenging. In Particular, automatic speech recognition (ASR) of natural audio streams is difficult and the resulting transcripts are not very accurate. The difficulty stems from a variety of factors including (i) noisy audio due to non-ideal recording conditions, (ii) foreign accents, (iii) spontaneous speech production, and (iv) diverse range of topics.

Our approach towards sentiment extraction uses two main systems, namely, automatic speech recognition (ASR) system and text-based sentiment extraction system. For text based sentiment extraction, we propose a new method that uses POS (part-of-speech) tagging to extract text features and Maximum Entropy modeling to predict the polarity of the sentiments (positive or negative) using the text features. An important feature of our method is the ability to identify the individual contributions of the text features towards sentiment estimation. This provides us with the capability of identifying key words/phrases within the video that carry important information. By indexing these key words/phrases, retrieval systems can enhance the ability of users to search for relevant information.

In this study, we evaluate the proposed sentiment estimation on both publically available text databases and YouTube videos. On the text datasets, the proposed system obtains ~95% accuracy on sentiment polarity detection (binary classification task) which is very competitive. On the YouTube videos, the proposed system obtains ~82% accuracy of sentiment polarity detection, which is very encouraging.

2. DATA COLLECTION

2.1. Database for Semantic-Sentiment estimation model

In order to train our text-based sentiment estimation system, we have used data from the following sources:

a) Amazon Product Reviews [9]
b) Pros and Cons database [10]
c) Comparative Sentence Set database [11]

The Amazon product reviews contain review comments about a large range of products including books, movies, electronic goods, apparel, etc. The Pros and Cons as well as the Comparative Sentence Set database contain a list of positive and negative sentiment words/phrases. From the combination of the three datasets, we extracted 800k reviews for training and 250k reviews for evaluation.

The review ratings for the Amazon dataset range from 1-to-5 stars. For this study, we convert the ratings into positive and negative classes where ratings above and below 2.5 are assigned to positive and negative sentiment, respectively. For the Pros and Cons dataset, and Comparative Sentence dataset, the comments were already labeled in a binary fashion.

2.2. YouTube audio database

YouTube videos [12] are an ideal choice for evaluation since they contain speakers using very natural and spontaneous speaking style. In order to establish ground truth on sentiment, three listeners viewed and rated the videos for sentiment. The listeners were asked to judge if the videos reflected positive, negative or neutral sentiment. Subsequently, the combined judgment of the listeners was used to select videos with positive and negative sentiments, and remove videos with neutral sentiment. In the end, we selected 28 videos (16 negative and 12 positive sentiment) containing expressive speakers sharing their opinion on a wide variety of topics including movies, products, and social issues. Our dataset contained 7 female and 19 male subjects. The average duration of these videos is 5 minutes, with individual videos ranging from 2 to 9 minutes. Three videos also had significant speech contribution from secondary speakers. The videos can be accessed using the following URL: http://bit.ly/YAgoYU

3. SENTIMENT MODEL DEVELOPMENT

In this study, we have used the text sentiment dataset described in Sec. 3.1 to develop the sentiment model. As shown in Fig 1, we first use POS tagging to generate a large set of initial text features that consist of nouns, adjectives, adverbs and verbs. We also extract text features that correspond to adjective-noun, verb-adjective, adverb-adjective and adverb-verb combinations. In the next step, we employ Maximum Entropy (ME) modeling technique to predict the ratings (positive and negative) given the text features extracted from review comments. This constitutes our baseline text-based sentiment model.

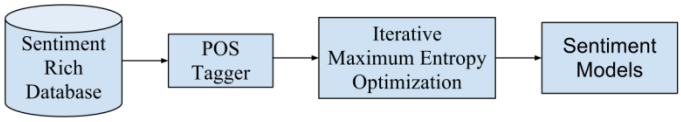

Fig 1: Sentiment model generation scheme.

One drawback of our baseline model is that it contains a large number of model parameters (~800k) since it follows a greedy training technique. Therefore, we propose an iterative training technique that can significantly reduce the number of model parameters while maintaining performance accuracy. The proposed iterative ME training process includes determining the most effective training features for every iteration and pruning the training feature list by eliminating ambiguous features (features that do not strongly predict positive or negative sentiment). The sentiment model training process is described in greater detail below.

3.1 POS Tagging

We initiate the process of feature extraction by performing part-of-speech tagging using the publicly available `Stanford's Log-linear POS Tagger [13][14]. We identify cluster combinations formed by combinations of nouns, verbs, adjectives and adverbs as important textual-features for sentiment prediction. Several studies have observed that noun tags in a review are likely to be product features, and adjectives capture the sentiments directed towards product features. Similarly, verbs and adverbs are likely to capture the product functionality and opinions. While extracting the initial feature-set based on POS-tags, it is ensured that adjective clusters, adverb-verb clusters, adverb-adjective clusters, verb-adjective clusters, verb-noun clusters, adjective-noun clusters, noun-adjective cluster combinations are extracted as features. Table 1 shows examples of word clusters extracted by using the mentioned technique.

Table 1: Example sentiment text-features extracted by using the POS-based approach.

POS Category	Examples	Sentiment
Adjective	Awful	Negative
Adverb-Adjective	really great	Positive
Noun-Adjective	distorted audio	Negative
Adjective-Verb	worth reading	Positive

3.2 Maximum Entropy Modeling

Using the sentiment text-features generated in the previous section, we develop our maximum entropy (ME) [14][15][16] based sentiment predictor.

Let y_j be the j^{th} sentiment where $y_j \in Y$ and $Y \equiv \{positive, negative\}$ is the set of all sentiment polarities. Let x_k be k^{th} review text, and f_i be the i^{th} text feature that captures the sentiment,

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

The functional definition above hypothesizes a relation between a feature present in the review text, and the corresponding review ratings. Using an evidence based modeling technique such as Maximum Entropy (ME), the relationship can be estimated quantitatively. The ME technique can predict the rating of the review y_j from its text 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 textual features f_i. The training data described in Sec 2.1 is used to develop the ME model for this study.

3.3 Maximum Entropy Model Tuning

Since the weights λ_{ij} of the ME model represent the importance of the feature, the ME technique essentially ranks the entire text based feature set. In fact, it is possible to segregate the feature set into mutually exclusive groups of ambiguous and unambiguous features. A feature is labeled as unambiguous when it is closely tied to a single sentiment alone, where an ambiguous feature is not strongly associated with a single sentiment,

$$x_k = \begin{cases} \text{ambigious} & \text{if entropy}\left(p(y|x_k)\right) \geq \tau, \\ \text{unambigious} & \text{if entropy}\left(p(y|x_k)\right) < \tau, \end{cases}$$

where τ is a threshold and $0 \leq \tau \leq 1$. It is useful to note that if $\tau = 0$, then the feature must be associated with a single sentiment alone, and this is the strictest condition.

Using the above principle allows us to prune the training feature set dramatically without significantly impacting performance. Starting with the baseline model obtained in Sec 3.2, we prune the training text-feature set by eliminating all text-features for which the entropy values lie above a given threshold. After the pruning operation, we train a new ME model using the pruned text-feature set. This process is repeated iteratively. It is noted that a number of stopping criterion can be employed for stopping the iterative training, namely, maximum number of iterations, desired accuracy or desired number of features (or model parameters).

4. AUTOMATIC SPEECH RECOGNITION

The speech recognition system used in this study was trained on a mixture of Switchboard and Fisher corpora (approximately 500 hours of audio data). We used MFCC (Mel-frequency cepstral coefficients) features for training the acoustic model. The acoustic model contained 4k tied-states and 16-mixture GMMs (Gaussian Mixture Models). A trigram language model was trained on a combination of Switchboard [17], Fisher [18] and UW191 [19] (191M words collected from the web by the University of Washington). The recognition lexicon contained 28K words.

Fig 2 shows our final system architecture. We use the mentioned ASR system to obtain text from YouTube video data. It is useful to note that 1-best hypothesis, lattices and nbest-lists can be used in the proposed system. However, in this study, we have only used the 1-best hypothesis. Once the decoded text for the video is obtained, we use the POS-tagger based feature extraction technique to identify useful sentiment features. Using these sentiment features, the ME-based sentiment model is used to estimate the sentiment polarity. The final output of our system are the probabilities of positive and negative sentiment (and they sum to unity).

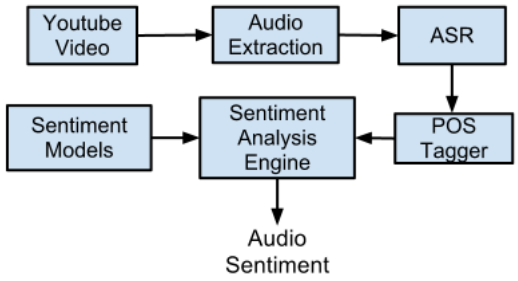

Fig 2: Proposed sentiment estimation system consists of an ASR system to convert speech to text, and a text-based sentiment estimation system to compute sentiment given the ASR decoded text.

5. RESULTS AND DISCUSSION

Table 2 shows performance of the proposed ME based sentiment estimation model on text data. It can be seen that the performance of the baseline ME model is 92.1%. Furthermore, the impact of ME model tuning is also seen as the system performance improved by 2.7% (94.8%) with a 32% feature set reduction. Additionally, we also trained a variation of the ME model which does not use noun based text features (Without-Noun system). The motivation behind removing noun features was to prepare a more domain-independent sentiment model, particularly for YouTube videos. Interestingly, this strategy provided improved results for text-based review data as well (2.2% absolute improvement over baseline).

Finally, it is also seen that ME model tuning further improves the without-noun system, and we obtain a final classification accuracy of 96.3% on text data.

Table 2: Binary sentiment classification accuracy on text data using baseline and without-noun systems.

	ME Model	ME Model with Tuning
Baseline	92.1%	94.8%
Without Nouns	94.3%	96.3%

In order to evaluate the YouTube videos, we decoded the audio streams using the ASR system described in Sec. 4. For this study, we only used the 1-best hypothesis for sentiment estimation. Using the decoded ASR text and the POS tagging system, we extract text-based sentiment features. Finally, the ME model is used to estimate the sentiment polarity using the extracted text features.

Table 3 provides the detection accuracy of the proposed system on YouTube data. Using the baseline ME model, we obtain an overall accuracy of 68% (75% and 62.5% detection accuracy for positive and negative sentiment, respectively). Once again, the impact of using the without noun system is seen as the overall accuracy improves to 82% (14% absolute improvement over baseline). It can also be seen that the performance on YouTube data is significantly lower than text data (82% vs. 96%). This difference can be attributed to noisy transcripts as a result of imperfect ASR due to spontaneous speech, accents and difficult recording conditions.

Table 3: Binary sentiment classification accuracy on YouTube data using baseline and without-noun systems.

ME Model with tuning	Accuracy		
	Positive Sentiment	Negative Sentiment	Overall
Baseline	75%	62.5%	68%
Without Noun	91.6%	75%	82%

As mentioned in Sec. 3.3, we can automatically extract the most unambiguous text features using the proposed ME system. This ability allows us to automatically extract key words/phrases for each YouTube video. In Table 4, we show example key words/phrases using this process. For example, the system automatically detected "incredible" and "better audio" in an Apple iPhone review video, and "violent" and "absolutely ridiculous" for a video on fighting in ice hockey.

It is worth mentioning that automatic extraction of sentiment words/phrases could be useful input for information retrieval systems.

Table 4: Sample word clusters picked by the system to derive Sentiment from the text obtained from ASR system.

Key words/phrases	Sentiment	Topic
Very good, amazing, Incredible, better audio	Positive	Apple iPhone review
Absolutely ridiculous, wrong, Violent	Negative	Fighting in Hockey
Pretty cool, affirmative, favorite	Positive	Transformer toy review
Disappointed, very exhausting, disastrous, very bad	Negative	Marijuana should not be legalized
Very comfortable, very durable, great	Positive	Skull Candy ear buds review
Totally absurd, really terrifying, really scary, crazy	Negative	Matt Damon's public comment about Sarah Palin

6. CONCLUSION

In this study, we have proposed a system for automatic sentiment detection for spontaneous natural speech and evaluated this on YouTube data. The proposed system uses ASR to obtained transcripts for the videos. Next, a sentiment detection system based on ME modeling and POS tagging is used to measure the sentiment of the transcript. We have also demonstrated ME model tuning and feature selection strategies that provide more accurate and domain-independent models. Our results show it is possible to automatically detect sentiment in natural spontaneous audio with good accuracy. Furthermore, we have also shown that our system is capable of providing key words/phrases that can be used as valuable tags for YouTube videos.

7. RELATION TO PRIOR WORK

Automatic sentiment extraction using text data is a very mature area of research [2-8]. This study extends sentiment extraction capability to natural instantaneous audio which has received little attention.

8. REFERENCES

[1]. Mishne and Glance, "Predicting movie sales from blogger sentiment," in AAAI 2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, 2006.

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

[3]. B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis." Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135, 2008.

[4]. B Pang and Lee "Foundations and Trends in Information Retrieval" 2(1-2), pp. 1–135, 2008.
[5]. 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.
[6]. N. Jindal, and B. Liu. “Opinion Spam and Analysis.” Proceedings of the ACM Conference on Web Search and Data Mining (WSDM), 2008.
[7]. Barbosa and Feng. “Robust sentiment detection on twitter from biased and noisy data”. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 36–44, 2010.
[8]. Gamon. “Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis”. Proceedings of the 20th International Conf on Computational Linguistics, 2004.
[9]. www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
[10]. Ganapathibhotla, Liu: “Mining Opinions in Comparative Sentences.” COLING, pages 241-248, 2008
[11]. Jindal, Liu: “Identifying comparative sentences in text documents.” SIGIR, pages 244-251, 2006.
[12]. www.youtube.com
[13]. Toutanova, Klein, Manning, and Singer, “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network”. In Proceedings of HLT-NAAC, pp. 252-259, 2003.
[14]. Ratnaparkhi, “A Maximum Entropy Model for Part-of-Speech Tagging”, In Proceedings of the Empirical Methods in Natural Language Processing, pp. 133-142, 1996
[15]. Csiszar. Maxent, “mathematics, and information theory. In K. Hanson and R. Silver, editors, Maximum Entropy and Bayesian Methods.” Kluwer Academic Publishers, 1996.
[16]. Rosenfeld, “Adaptive Statistical Language Modeling: A Maximum Entropy Approach.” PhD thesis, Carnegie Mellon University, 1994.
[17]. Godfrey, Holliman, McDaniel, “SWITCHBOARD: Telephone speech corpus for research and development,” Proc, IEEE ICASSP, pp. 517-520, 1992
[18]. Cieri, Miller, Walker, “The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text.”, LREC 2004
[19].http://ssli.ee.washington.edu/ssli/projects/ears/WebData/web_data_collection.html