This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2678164, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Automatic Sentiment Detection in Naturalistic Audio

Lakshmish Kaushik, Abhijeet Sangwan and John H. L. Hansen
Center for Robust Speech Systems (CRSS), Eric Jonsson School of Engineering and Computer Science,
The University of Texas at Dallas, Richardson, Texas, U.S.A.

***Abstract*—Audio sentiment analysis using automatic speech recognition is an emerging research area where opinion or sentiment exhibited by a speaker is detected from natural audio. It is relatively under-explored when compared to text based sentiment detection. Extracting speaker sentiment from natural audio sources is a challenging problem. Generic methods for sentiment extraction generally use transcripts from a speech recognition system, and process the transcript using text-based sentiment classifiers. In this study, we show that this baseline system is sub-optimal for audio sentiment extraction. Alternatively, new architecture using keyword spotting (KWS) is proposed for sentiment detection. In the new architecture, a text-based sentiment classifier is utilized to automatically determine the most useful and discriminative sentiment-bearing keyword terms, which are then used as a term list for KWS. In order to obtain a compact yet discriminative sentiment term list, iterative feature optimization for maximum entropy sentiment model is proposed to reduce model complexity while maintaining effective classification accuracy. A new hybrid ME-KWS joint scoring methodology is developed to model both text and audio based parameters in a single integrated formulation. For evaluation, two new databases are developed for audio based sentiment detection, namely, YouTube sentiment database and another newly developed corpus called UT-Opinion Opinion audio archive. These databases contain naturalistic opinionated audio collected in real-world conditions. The proposed solution is evaluated on audio obtained from videos in youtube.com and UT-Opinion corpus. Our experimental results show that the proposed KWS based system significantly outperforms the traditional ASR architecture in detecting sentiment for challenging practical tasks.**

***Index Terms*—Sentiment Analysis, LVCSR, Keyword spotting, Maximum Entropy, Opinion, Amazon, UT-Sentiment Audio Archive**

I. Introduction

TExt based sentiment detection is an established field in natural language processing (NLP). Sentiment analysis / opinion mining, analyzes peoples opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [23]. There is an enormous amount of opinionated data in the social media and on the Web in the form of Twitter, Facebook, message boards, blogs, and user forums [1], [2], [23], [6], [44], [5]. The decision making process of people is affected by the opinions formed by a wide range of thought leaders and ordinary people over the web. Amazon, Yahoo, Google and various other personalized websites are a significant resource for obtaining opinions concerning products of any kind. Many consumers form their decision to buy a product dependent on feedback from online reviews. This information not only helps ordinary people make decisions, but also provides indicators for companies about the reception of a product, or a political context, to understand the mood of people regarding an ongoing social/cultural/political/economic issue. Typically, a given text is classified to exhibit positive, negative or neutral sentiment [6], [9], [10], [11], [22], [23], [24], [25], [26], [27], [29], [30], [36], [37]. This form of automatic classification has numerous applications such as measuring public opinion/sentiment using Twitter feed, analyzing online product reviews, understand mass social human behavior over a topic, product or an event [7], [8], [15], [16], [17], [18]

Text based reviews form only one of the many ways people can express their sentiment/opinion about products or social issues. Audio/Video is also a prominent method to express opinions. Millions of videos on YouTube [65] are about products and movie reviews, product un-boxing, political, social issue analysis and opinions on them. There are many audio platforms on the Internet where individuals express their opinions. Also, the audio mode is more powerful than text for many situations because they provide richer cues of the speaker regarding their opinions. This vast resource is untapped and extracting sentiment/opinion of society about specific products or mass opinion regarding social or political situations will be very useful for information analysis. Detecting sentiment in audio is still an unexplored area.

Speech based sentiment extraction is an emerging and challenging field. In this study, robust methods are presented to extract sentiment/opinion from natural audio sources. A hybrid system is developed which utilizes a robust Automatic Speech Recognition (ASR) system in tandem with NLP based sentiment analysis techniques to detect sentiment of audio streams. Unlike text based sources, audio sources have a high degree of variability both in terms of expressing opinion as well as the mode of expression of the opinion. There are a range of challenges for sentiment extraction in highly natural speech sources including:

1) *Domain and vocabulary* : The speaker can express opinions about any topic, (e.g., products, movies, politics, social issues, games, etc.) Hence the ASR system should be efficient to handle a wide range of domains and vo-

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. Corresponding author: Lakshmish Kaushik (email: lakshmish.kaushik@utdallas.edu).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2678164, IEEE/ACM Transactions on Audio, Speech, and Language Processing

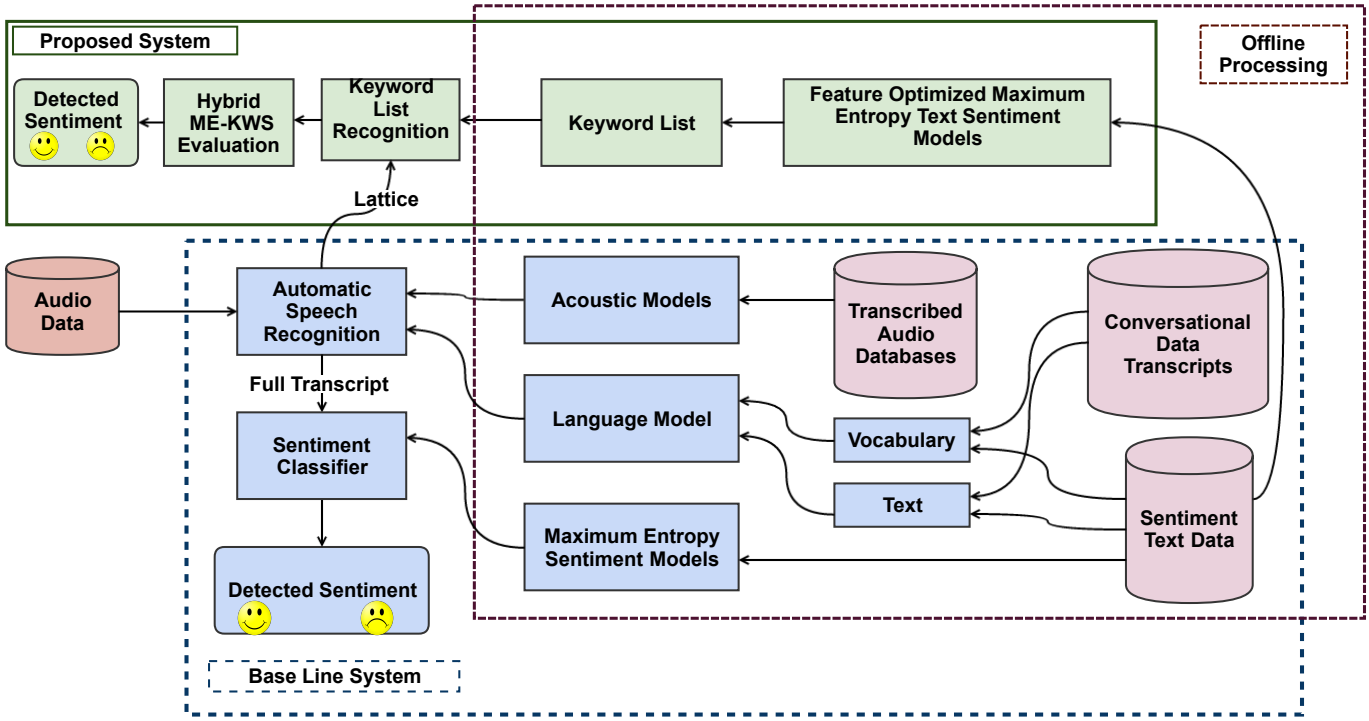

Fig. 1. Baseline audio sentiment detection system uses Automatic Speech Recognition (ASR) to generate transcripts which are then processed by the sentiment classifier. Proposed sentiment detection system uses Keyword Spotting (KWS) to detect sentiment-bearing terms in the audio data. An iterative feature pruning technique is proposed to generate the sentiment-bearing terms for KWS. Offline development of customized text based sentiment models.

cabulary. The language model should be comprehensive.

2) *Speaker variability and speaker accents* : ASR system should be robust to speaker variability which includes a wide range of English accents from all over the world.
3) *Noisy audio and channels* : Inconsistent recording equipment and different mode/distance of recording, inconsistent acoustic and background environment conditions make the sentiment detection problem challenging. Also, background music/talk, intentional music mixing, reverberation issues make the problem harder.
4) *Natural and Spontaneous* : Detecting audio sentiment in natural and spontaneous speaker settings and various speaker interactive scenarios (i.e., 1-way, 2-way, public speech etc.) is challenging.

Given the explosive increase of online videos on product reviews, un-boxing, politics, sports, culture, *etc.* on websites such as YouTube.com, Vimeo, News broadcasting, Daily Motion, Twitch and Vine, automatic audio sentiment detection technology would be useful in collecting and summarizing information for users.

Figure 1, shows all various aspects of the baseline and proposed audio based sentiment detection. There are three core subsystems. First one is the offline text based sentiment model generation, the second is the ASR based sentient detection system forming our baseline system, and finally the third is a proposed system using audio Keyword Spotting (KWS) approach. Each block is explained in detail in subsequent sections. Robust sentiment model development is explained in Section III. Next, many strategies for audio sentiment detection using speech recognition are presented. In Section IV, a continuous speech recognition system is developed for this purpose. Advantages and disadvantages of the formulated preliminary sub-tasks based on the baseline ASR transcript have previously been explored in tandem speech sentiment system discussed in [12], [13]. In reality, accurate sentiment detection generally relies on a small fraction of the speech recognition transcript, because sentiment bearing vocabulary tends to be sparse in spoken opinions. For example, in a statement like "I watched Game of Thrones last night and it was wonderful", only 1 word out of 11 conveys the some sort of sentiment. While this may not be true for every utterance sequence, sparseness is generally prevalent in spoken comments. Given this nature of spoken comments, it would be reasonable to assume that sentiment detection is tolerant of moderate word error rates (WERs), and this is precisely what we observed in our previous study [13]. In other words, sentiment detection accuracy depends on being able to reliably detect and recognize a very focused vocabulary in the spoken comment audio stream. Therefore, keyword spotting (KWS) technology is expected to be better suited for sentiment detection, as opposed to full-transcript ASR. In this study, a detailed analysis and KWS sentiment system development is presented in Section V. A new hybrid ME-KWS audio sentiment modeling method is also proposed. In this study, the proposed system is evaluated on two different corpora: UT-Opinion and videos from YouTube.com. The UT-Opinion (Section VII-B) is a new corpus specifically collected for the purpose of audio sentiment detection. The corpus contains interview style data, where subjects give their opinion on various topics in natural settings around a university campus.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2678164, IEEE/ACM Transactions on Audio, Speech, and Language Processing

YouTube database (Section VII-A) is a collection of videos from youtube.com which was carefully chosen for the purpose of sentiment research. Our experimental results show that the proposed KWS framework significantly outperforms the conventional ASR approach on both datasets. Results and analysis are presented in Section VIII.

II. Background

Sentiment analysis can be classified mainly into the following four categories,

(a) Document-level sentiment analysis: The sentiment is generated on the overall document/review level. This is a global analysis. References include: [45], [42], [39], [46];

(b) Sentence-level sentiment analysis: This gives a microlevel sentiment assessment for every sentence. This is effectively a local analysis. References include: [47], [48], [49]

(c) Aspect-based sentiment analysis: This gives a sentiment variation in both a local and global level. Aspect-based sentiment analysis focuses on the recognition of all sentiment expressions within a given document, and the aspects/objects to which they refer [41], [51]. It is possible to have varying local sentiment with one overall document assessed value.

(d) Comparative sentiment analysis: Extracting opinion in reviews where a product is compared with other product [60] (e.g.,: Google search is better than Yahoo).

The first step in development is to pre-process text using linguistic tools such as stemming, tokenization, parts-of-speech tagging, entity extraction, and relation extraction depending on classification technique [63], [51], [14]. Once we have the features generated, next comes classification. There are two main approaches to document level sentiment analysis: supervised learning and unsupervised learning. The supervised approach assumes there is a finite set of classes into which the document should be classified, and training data exists for each class including positive and negative, with neutral being an option (i.e., a 2-class or 3-class system). Using various supervised methods such as Maximum Entropy [21], SVM [40], Naive Bayes [40], Logistic Regression, or KNN, features can be learned to classify the sentiment exhibited. There are also unsupervised approaches to document-level sentiment analysis using semantic orientation of specific phrases within the document. Most methods use supervised approaches to classify the sentences into the two classes. The above techniques are used for document level classification. There are very new methods using Neural Networks [52], [54] and Recursive Neural Tensor Network (RNTN) [49] which analyze sentiment at the sentence level. Of all these methods, we use the ME based document level sentiment detection strategy in this research.

III. Maximum Entropy Text Sentiment Detection

Maximum Entropy (ME) is a multinomial logistic regression method that predicts the probabilities of different possible outcomes of a categorically distributed dependent variable, given a set of independent variables. Maximum Entropy models offer a way to combine diverse pieces of contextual evidence in order to estimate the probability of a certain linguistic class occurring with a certain linguistic context. This is achieved by estimating the probability of class a occurring with context b [21]. To develop the ME sentiment models, relevant features that reflect both positive or negative sentiment from the text are extracted and trained.

Maximum entropy based sentiment system is an effective approach for discriminative learning of features. This can be used in huge database scenarios to develop probabilistic models that can predict sentiment effectively. Sentiment analysis typically predicts positive, negative or neutral sentiment given a sentence. In this study, we consider sentiment classification as a two-class problem, (*i.e.*, Positive *vs.* Negative classification), and the Maximum Entropy based approach is used to develop the sentiment classifier. Figure 2 shows the methodology used to develop the text sentiment models. In what follows, each step is explained in detail.

A. Sentiment feature representation and generation

Table I shows a sample product review from the website amazon.com, along with the corresponding star rating. The star rating is an objective measurement of sentiment on a 1-to-5 point scale, where greater and lesser number of stars correspond to positive and negative sentiment, respectively. It is common practice to use a star rating as ground truth for reviews. In this study, reviews with greater and lesser than 3 stars are assumed to convey positive and negative sentiment, respectively. Comments with 3 star ratings are discarded as they tend to be a mixed collection of both positive and negative comments, and can be disruptive for training.

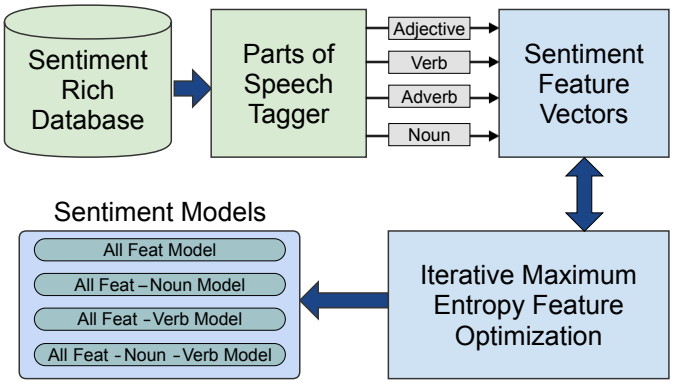

Fig. 2. Steps for generating sentiment models using the proposed Iterative Maximum Entropy Optimization algorithm. Using a sentiment rich text database, parts-of-speech tags (*i.e.*, Adjective, Verb, Adverb, Nouns combinations) are used to extract sentiment features. Iterative Maximum Entropy Optimization is used to learn sentiment classification models. In this study, we generate four variations of sentiment classification models.

In the first step towards feature extraction, we extract words and word combinations which potentially depict sentiment. This is done by parsing the raw text using a part-of-speech (POS) tagger. As shown in Figure 2, we use Stanford's Log-linear POS tagger system [62], [63]. After tagging, words and word combinations formed by particular combinations of adjective (JJ), Verb (V*), Adverb (RB*) and Noun (N) are extracted as sentiment bearing features. In particular, the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2678164, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE I
EXAMPLES OF RAW INPUT FROM AMAZON REVIEWS AND THEIR NORMALIZED EQUIVALENTS. STAR RATING IS BINARIZED DEPENDING ON THE STAR VALUE (RANGE: 0-5) WHICH ACTS AS GROUND TRUTH. SENTIMENT FEATURE VECTOR IS EXTRACTED FROM EACH REVIEW USING POS TAGGER FOLLOWED BY SENTIMENT MODEL TRAINING.

Star Rating	Binary Rating	Original Review	Extracted Sentiment Features
★★★★★	1	The Phones **Works Excellent**, **great speed**, the cam get really **great quality**. Android+HTC sense is a **nice interface**.	excellent, great_speed, great_quality, nice_interface
★★★★★	0	**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 Battlefield Earth than subject myself this **nonsensical rambling**.	Horrible, Garbage, worst_pieces, Thoroughly_discourage, Nonsensical_rambling

following regular expressions are used: (i) JJ, (ii) JJ-JJ, (iii) JJ-V*, (iv) V-JJ, (v) RB-JJ, (vi) RB-V, (vii) V*-RB, (viii) JJ-NN, (ix) NN-JJ, and (x) V*-NN. Table I shows this process of extracting sentiment bearing features for the sample text. The automatically identified features are shown in bold face. In this study, a large dataset derived from various online sources is used to extract the mentioned sentiment bearing features (more details about the dataset can be found in Section VI). In the next section, we explain how the ME classifier is trained using sentiment bearing features.

B. Maximum Entropy Classifier

We use Maximum Entropy to determine the comment sentiment polarity given the textual sentiment features as input [19], [20], [29]. Let y_j be the sentiment where $y_j \in Y$ and $Y \equiv \{positive, negative\}$ is the set of sentiment polarities. In this case j has two values (positive and negative). Let x_k be k^{th} textual sentiment feature, then function f_i is defined as:

$$f_i(x_k, y_j) = \begin{cases} 1 & \text{If } x_k \text{ is present in text comment,} \\ 0 & \text{Otherwise.} \end{cases} \quad (1)$$

Equation 1 hypothesizes the relation between a feature present in the review text, and the corresponding review rating. Sentiment features from each review with the corresponding sentiment truth (1 or 0) is applied as input to the classifier which learns over many instances of the feature. Applying an evidence based modeling technique such as 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:

$$p(y_j|x_k) = \left\{ \frac{1}{Z_\lambda(x)} \sum_{i=0}^{N} \exp(\lambda ij f_i(x_k, y_j)) \right. \quad (2)$$

where, $Z_\lambda(x)$ is a normalizing term, and λ_{ij} are weights assigned to the function f_i.

C. Feature Selection

All feature combinations noted in the previous section do not contribute to the sentiment accuracy equally. In order to better understand the relative strengths of feature combinations, we performed experiments where four different groups of features were used to build sentiment detection models. In particular, the following groups were used: (a) All POS tags combinations are employed as features, (b) All POS tag combinations except features that use noun POS tags, (c) All POS tag combinations except features that use noun and verb POS tags, and (d) All POS tag combinations except features that use verb POS tags. Four different sentiment classification models were built using these four feature groups. By comparing the sentiment accuracy of these models using a common evaluation set, it was found that the system that uses feature group (b) provided the highest accuracy, and the system that uses feature group (a) provided the least accurate model. This suggests that noun-based features can potentially reduce sentiment detection accuracy, and therefore we use feature group (b) for the remainder of this study. To see further experimental details, interested readers can review our previous work [13].

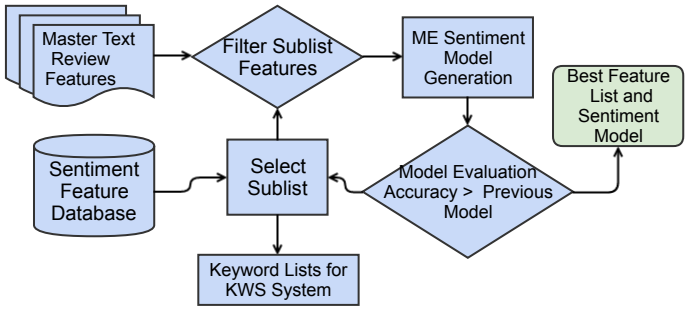

Fig. 3. Steps for the proposed Iterative Maximum Entropy Feature Optimization. It generates the a highly optimized features and an optimal text sentiment model. This method generates a highly discriminative set of sentiment keywords.

D. Iterative Maximum Entropy Feature Optimization

Even after removing noun based features (as described in the previous section), the sentiment bearing feature set is rather large, (*i.e.*, we obtain more than 322,000 sentiment features on our datasets). Such a large feature set can be somewhat redundant for audio based sentiment recognition, as automatic speech recognition solutions typically use relatively

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2678164, IEEE/ACM Transactions on Audio, Speech, and Language Processing

smaller vocabulary. In other words, the output transcript (even if we were to obtain 100% accurate transcripts) would never contain certain features due to restricted ASR vocabulary. Experimental tests on text based sentiment detection show that reducing the size of the feature set adversely impacts accuracy. Therefore, there is a need to reduce the size of the feature set in a manner such that the most discriminative and effective features are retained, thereby minimizing the impact of accuracy of the sentiment detection models. This motivates the development of an Iterative Maximum Entropy Feature Optimization, where we develop a new iterative method to reduce feature set size while retaining the most effective features for sentiment detection.

Features with lower frequency may give a good classification probability in MaxEnt, but being rare it doesn't contribute significantly to the model sentiment like high frequency features. Pruning higher frequency features will be costlier than pruning features with lower frequency. Previously published Iterative Maximum Entropy Optimization method in [53], prunes the features by varying MaxEnt probability threshold margins in each iteration. By doing so many high frequency features were relatively nearer to equi-probable point than rarely occurring features were pruned (they occurred in both positive and negative reviews in large numbers). Hence, to avoid high frequency feature loss which could contribute to develop an efficient discriminative sentiment model a new iterative pruning method based on feature frequency is proposed.

First, all 322,000 sentiment features are rank-ordered by their frequency of occurrence using a dataset of 72 million reviews obtained by CRSS-UTDallas from all the sources mentioned in Section VI. Let N_i be the cutoff for the i^{th} iteration, where $N_i > 0$. Features that have frequency of occurrence greater than or equal to N_i are chosen for training the sentiment model M_i. We increase the cutoff value by 25 for every iteration, (*i.e.*, $N_{i+1} = N_i + 25$, where $N_1 = 0$). Therefore, all the features are used for training in the first iteration and the size of the sentiment detection model decreases with every iteration, as we discard more features from training. Since the original feature is rank ordered by frequency of occurrence, we start with removing the most rarely occurring features first. Hence, we expect to see minimum impact on accuracy along with significant reduction in overall feature set size in the initial iterations. Subsequently, with more iterations, we expect to see an increasing impact on accuracy as more low quality features are removed from training. For speech based sentiment detection, we want to choose the sentiment model that balances feature set size with accuracy.

Figure 4 shows model accuracy *vs.* feature set size along with number of iterations (i) and corresponding feature cutoff (N_i). The training starts with a minimum cutoff frequency of 0 and was increased up-to 1300 with a step size of 25 (as explained previously). It can be seen from the figure that the feature set size reduces significantly from about 322K to 60K (absolute reduction of 262K) during the first few iterations ($i \leq 5$), with accuracy decreasing only from 91.7% to 91.5% (absolute reduction of 0.2%). In other words, the loss in accuracy is marginal but the reduction in feature set is significant. In subsequent iterations ($i > 5$), the accuracy falls much more sharply, and the feature set size reduction is relatively smaller. For example, for iteration $i = 53$, the model accuracy is 90.6% (about 1.1% absolute reduction), with a feature set size of about 10K (an absolute reduction of about 312K). When compared to the previously proposed probability margin based pruned sentiment model the new sentiment model developed using frequency based pruning saw a 1.4% relative increase in text based sentiment classification accuracy.

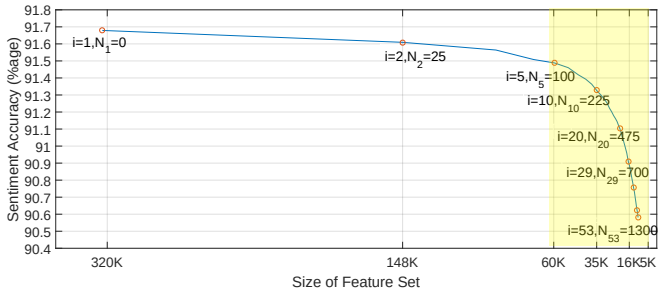

Fig. 4. Feature set reduction using the proposed Iterative Maximum Entropy Feature Optimization algorithm. The algorithm prunes the least effective features in the initial iterations and provides significant reduction in feature set size with limited impact on sentiment detection accuracy. The models in the shaded area are practical for most automatic speech recognition systems.

In this manner, the proposed Iterative Maximum Entropy Feature Optimization algorithm provides smaller models that are more easily used with ASR with only a minimal impact on sentiment detection accuracy. In fact, models under 60K as a feature set size are usable with speech recognizers as the feature set vocabulary can be absorbed into the ASR vocabulary while keeping the ASR vocabulary size manageable. In this study, we use feature sets with sizes below 64K for audio sentiment recognition experiments.

IV. Baseline Sentiment Detection for Audio

A simple way to perform sentiment detection for audio data is to process ASR output transcripts through a text-based sentiment detection system (such as the ME method described in previous sections). In this study, this system constitutes the baseline for audio based sentiment detection experiments.

Figure 1 gives an overview of how ASR based sentiment extraction works. The automatic speech recognition (ASR) system used in this study is the KALDI speech recognition system [72]. The acoustic models are trained using a mix-style approach, where acoustic data from multiple corpora are used. We have used Conversational Telephone Corpora consisting of Switchboard and Fisher corpora to train our acoustic models. The above corpora constitutes around 2000 hours of speech data. Part of the training data is corrupted with additive noise at various SNRs (signal to noise ratio) to create a mixture of clean and noisy conditions (to mimic various kinds of noises like office, fan, cafeteria, car road which we expect to see in the evaluation corpora used in this study). The acoustic models use Mel frequency cepstral coefficients (MFCCs) along with deep neural networks (DNNs). The ASR vocabulary size is about 120K words. The vocabulary contained in the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2678164, IEEE/ACM Transactions on Audio, Speech, and Language Processing

64K sentiment feature set (described in previous sections) is part of the ASR vocabulary. Additionally, the language model for the ASR is trained on data from two sources, namely, CTS (conversational telephone speech) corpora and reviews dataset (described in Section VI). Triphone language models are trained using various text sources like Switchboard [73], Fisher [74] and UW191 [75] (191M) words collected from the web by the University of Washington) and entire six sentiment corpora explained in section VI. Altogether, the combined text dataset for language model training contained more than 1 billion words. In this manner, the language context of the sentiment features is captured in the language model.

The ASR uses a 2-pass decoding strategy, where an fMLLR transform is estimated in the first pass and used to normalize the MFCCs. In the second pass, the normalized MFCCs are used with DNN models to obtain the output recognition lattice. Subsequently, one-best transcripts are obtained from the lattice. Finally, the one-best transcripts are analyzed by the Maximum Entropy sentiment detection model described in Section III-D to obtain the sentiment decision for the audio file.

V. Proposed Keyword Spotting Sentiment Detection System for Audio

Sentiment bearing terms are relatively sparse in both speech and text. Therefore, a large majority of the ASR transcript, (as well as the errors in the transcript), have limited impact on sentiment detection accuracy. When compared to ASR, Keyword Spotting (KWS) provides the ability to focus on relevant terms and phrases, and ignore irrelevant (from an application perspective) background text. Therefore, it is suggested that KWS would be better suited for sentiment detection as opposed to employing full ASR transcripts directly.

In order to further explain the above mentioned argument, we present a simple experiment that demonstrates the impact of ASR WER (word error rate) on sentiment detection. In this experiment, we add controlled amounts of substitution errors (to simulate a required word error rate) in our text based evaluation dataset. While simulating substitution errors, words were randomly chosen and replaced by other words from the ASR vocabulary. This process generates various noisy versions of the evaluation dataset. Using our best sentiment detection model, we then compute sentiment detection accuracy for these noisy variants, and compare this accuracy to the original clean version.

Figure 5 shows sentiment detection accuracy obtained on noisy variants of the evaluation dataset *vs.* the amount of controlled WER introduced in the transcript. This shows the tolerance of sentiment detection to ASR word errors. It can be seen that as simulated WER increases from 0% to 80%, there is a gradual, almost linear drop in average sentiment detection accuracy from 91.7% to 80%. When simulated WER increases beyond 80%, we see a sharp decrease in accuracy (and the system reaches random performance, (*i.e.*, 50%) at 100% simulated WER). This simple experiment shows that automatic sentiment detection in audio can tolerate high amounts of WER. In other words, successful sentiment detection depends on finding the few effecting sentiment relaying terms/phrases. This finding is further strengthened by empirical observations that sentiment bearing terms are relatively rare in text and audio (for example, approximately 2-to-5% relative frequency of occurrence in reviews). Taken collectively, this implies that sentiment detection should be possible with KWS using a relatively small keyword list, and it is likely to outperform full transcript ASR. Though an ASR system might generate word errors differently than a random substitution word error simulation; however the sentiment distance between the correct ASR word errors vs Randomly selected word errors should be equally large/divergent. Therefore we feel the simulation accurately represents the system's ability to accurately access sentiment of the content. It is difficult to force a predetermined WER in ASR systems for Keyword Spotting.

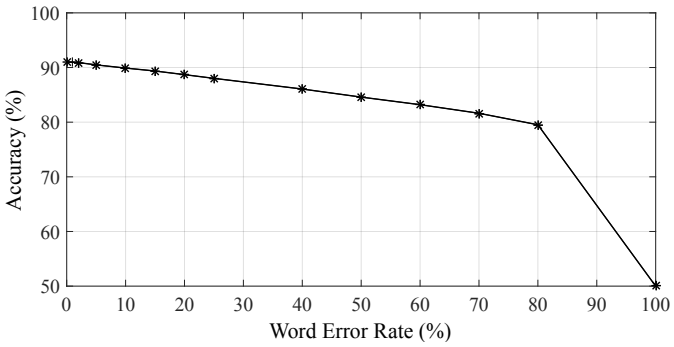

Fig. 5. A plot of the variation in sentiment detection accuracy *vs.* simulated word error rate (WER). The trend suggests that only a small fraction of words can effectively determine sentiment, and most words are irrelevant from a detection perspective. This implies that KWS may be a feasible and superior solution for sentiment detection over ASR.

A. Keyword spotting based sentiment detection system

In order to build an effective KWS system, we need a compact yet effective keyword list. The Iterative Maximum Entropy Feature Optimization algorithm described in Section III-D allows us to extract the most effective and compact feature sets. As previously shown in Figure 3, each iteration provides an increasingly smaller feature set (which can be directly used as the keyword list). Using this approach, we create keyword lists of small (approximately 200, 500, 1K, and 2K), to medium (approximately 4K, 8K, and 16K), and large (approximately 32K, 48K, 64K and 80K) word count sizes.

B. Keyword Spotting System

The KWS system developed is an extension of our KALDI based large vocabulary ASR system from Section Section IV. Using the word lattice generated by the ASR previously in Section IV, a finite state transducer (FST) based method is used to search the word lattices for keywords [72]. In parallel, the word lattices are converted in to phone lattices, and the PCN-KWS (phone confusion network keyword spotting) method employed to search for keywords as well [70], [71]. Subsequently, the search results from the two methods are combined (by simple likelihood combination) to yield the final keyword result list. The KWS system assigns log-likelihood

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2678164, IEEE/ACM Transactions on Audio, Speech, and Language Processing

values to keyword hypotheses in the audio. The higher the value of the log-likelihood, the more confident the hypothesis. Keywords with a likelihood below a certain pruning threshold are rejected, and not used for further computation. In our experiments, we have rejected keywords with log-likelihood below -3 (this threshold value is empirically determined and yields a good equal error rate across many KWS experiments).

C. KWS-ME Hybrid Evaluation System

Maximum entropy is used to compute the sentiment detection likelihood using Eq. 2. The function $f_i(x_k, y_j)$ in this equation assumes binary values of 0 and 1 given either the absence or presence of evidence (*i.e.*, sentiment bearing feature), respectively. Unlike text, using KWS for detecting sentiment bearing terms in audio produces the likelihood of the sentiment bearing feature (a probability number that is between 0 and 1). In order to incorporate keyword detection confidence into the sentiment detection likelihood computation, we propose to set $f_i(x_k, y_j)$ equal to the exponent of the log-likelihood. In this manner, $f_i(x_k, y_j)$ can now assume values between 0 and 1, and it represents the confidence of detecting the sentiment bearing feature. The benefit of computing $f_i(x_k, y_j)$ is this manner is that features that are detected with higher confidence have a greater say in determining the overall sentiment detection outcome. In other words, let c_k be the confidence score of k^{th} detected sentiment bearing feature. Now, $f_i(x_k, y_j)$ in Eq. 2 is computed as,

$$f_i(x_k, y_j) = c_k, \text{where } 0 \leq c_k \leq 1 \tag{3}$$

Using the above equation, we compute a new value of the ME sentiment detection likelihood which now incorporates keyword detection confidence into scoring.

VI. Text databases

In this study, we have used the following text corpora for our experiments:

1) Amazon Product Reviews [55], [57], [58],
2) Pros & Cons database [59],
3) Comparative Sentence Set Database [60],
4) Opinion Lexicon Database [55],
5) Scale database [14], and
6) R-T polarity database [55].

The noted selection attempts to obtain data from various domains (ensuring diversity). Data from such varied sources tend to have inconsistencies in the rating and data representation. For example, some reviews are graded on a rating scale of 1-10, and others on 1-5. In this study, the star ratings for all reviews are normalized by first transforming all ratings to a 5-point scale. Subsequently, ratings of 3 are discarded, ratings below and equal to 2 are labeled negative, while ratings above and equal to 4 are rated positive.

The total dataset contains approximately 53 million reviews, 43 million positive, and 10 million negative. This dataset is used to create balanced training and evaluation datasets for our experiments. In particular, 8 million reviews (4 million positive and 4 million negative reviews) are used for training, 1.1 million reviews for evaluation ($550K$ positive and $550K$ negative), and 2 million reviews for development (1 million positive and 1 million negative). All datasets are chosen randomly and have no overlap.

VII. Audio Databases

Unlike text sentiment detection which has numerous data sources, audio corpora resources dedicated specifically for sentiment detection task are very limited [64]. In the proposed work we have developed databases in a very controlled manner keeping audio sentiment detection as the main focus. In this context, we have developed two new audio datasets to support our sentiment detection experiments.

A. YouTube Sentiment Audio Database

YouTube videos [65] are an ideal source to collect speech data where 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. The dataset contains a balanced speaker distribution. It has 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.

In order to establish the sentiment ground truth on the YouTube videos, three listeners evaluated the videos and rated them for positive and negative sentiment. The listeners were asked to judge the videos for positive or negative sentiment.

The evaluators were asked to watch the video and listen to the speaker while wearing headsets. They were asked to set volume to a comfortable level and the experiment was conducted in an office environment. Each evaluator judged the sentiment of the speaker in each video on a 5-point scale, where, 1, 2, 3, 4, and 5 meant extremely negative, negative, neutral, positive, and extremely positive, respectively. In order to determine the final ground truth, the ratings of the three evaluators were averaged to obtain one rating per video. Finally, average ratings of 1 and 2 were considered negative, 4 and 5 positive, and 3 neutral. These decisions were used as ground truth for the experiments presented in this study.

B. UT-Opinion Audio Archive

The YouTube dataset mentioned in the previous section has some drawbacks. It is hard to control factors such as age, gender and accent (and these are relevant variants towards measuring performance of a speech based system). Additionally, the sentiment expressed in YouTube videos is less likely to be subtle (*i.e.*, speakers tended to take extreme positions). Finally, it is hard to get a diverse set of opinions on the same topic (*i.e.*, we find it hard to control for topic as a variable in the study).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2678164, IEEE/ACM Transactions on Audio, Speech, and Language Processing

In order to overcome the above mentioned issues, we have developed the UT-Dallas Sentiment (UT-Opinion) Audio Archive. In this corpus, each participant was asked 10 questions to illicit opinion. The questions are:

1) Are you aware of our college Mascot TEMOC? Do you like it?
2) Which is the last movie that you saw? Did you like it?
3) Do you remember your last vacation? How was your experience?
4) Are you satisfied with the library and other campus facilities?
5) Are you satisfied with the university's bus/transport facilities?
6) Do you like living in Dallas?
7) What is your opinion about Google Glass?
8) Do you remember the last time you went for a fancy restaurant dinner? What was your experience?
9) Recall the last course you completed. Did you like the class?
10) Would you like a dog as your pet?

The questions were asked in an interview style, with the interviewer and participant facing each other. The responses were recorded using a Samsung hand-held tablet. The distance between the participant and the tablet microphone was between 1-to-2.5 ft. After that participant answered the question subjectively, they were also asked to rate their opinion on a scale of 1-to-5 (where, 1, 2, 3, 4, and 5: meant extremely negative, negative, neutral, positive, and extremely positive, respectively). This data was collected in various acoustic settings inside the campus, namely, classrooms, hallways, offices, library, gym and street. Altogether, 63 people participated in the study, 43 male and 20 female. Students, staff and faculty participated in the data collection. The ages varied from 19 to 62 years old. The corpus consists of a diverse speaker set including native speakers of American English as well as Arabic, Spanish, Farsi, Mandarin, Korean, Hindi, Tamil, Kannada, Marathi, Bengali, Telugu, Italian and French.

Using the method described in Section VII-A, 5 evaluators listened to the audio data and generated the sentiment ground truth. From the tabulated results, the Inter-Rater agreement for YouTube database is 94.5% and 80.1% for UT-Opinion dataset. It can be observed that when compared to YouTube dataset UT-Opinion has a lower agreement rate. This shows that the sentiment detection for UT-Opinion database is relatively not as easy when compared to YouTube database for human annotators. This is an indication that the sentiment detection for UT-Opinion dataset could also be a challenging task for machines.

VIII. Results and Discussion

In this section, experiments on text and audio based sentiment detection are presented.

A. Text Sentiment Results

In this experiment, we show the accuracy of the automatically generated feature set developed in Section III-C on the text evaluation corpus described in Section VI. The automatically generated feature set in compared to a handcrafted list of 6,800 sentiment features developed by researchers at the University of Chicago for the purposes of text sentiment detection [59]. The handcrafted feature set has been popular and used for building many text based sentiment detection systems [28], [33], [34]. The datasets mentioned in Section VI are used for this experiment.

TABLE II
Equal Error Rate for Automatic Feature Generated and Handcrafted Feature systems.

Dataset	Equal Error Rate (Error Reduction)
Hand Crafted Features	13%
Unigram and Bigram Features	8% (38%)

Table II shows the The equal error rate (EER) of the above mentioned systems. The EERs for automatically extracted feature set and handcrafted feature set are 8% and 13%, respectively. As expected, the automatically extracted feature set outperforms the handcrafted feature set. There is a 38% reduction in error when both Unigram and Bigram features are used. An interesting point to note is that the handcrafted feature set (7K features) is only 5% lower in accuracy compared to the automatically extracted feature set (322K). This shows that typically a small highly discriminative feature set determines the majority of the outcome in sentiment detection. From an audio based sentiment detection perspective, this is an important point since smaller keyword lists are more practical than larger ones.

TABLE III
Text based sentiment model results on comparative dataset

Datasets	Precision		Recall		F-Score	
	Naive Baye's	Max Entropy	Naive Baye's	Max Entropy	Naive Baye's	Max Entropy
Reviews	84%	**92%**	80%	**90%**	82	**91**
Articles	75%	**88%**	80%	**84%**	77	**86**
Forums	73%	**86%**	83%	**86%**	78	**86**

In another experiment, the proposed ME based sentiment system (with the proposed Iterative Maximum Entropy Feature Optimization) is compared to the Naive Baye's method for sentiment detection in Table III. The same feature set is used for both MaxEnt and Naive Bayes. In this experiment, both methods are evaluated on the widely used comparative dataset corpus which has 3 subdivisions, namely, product reviews, forum reviews, and articles [60]. The results are shown in Table III. It can be seen that the F-Score of proposed ME sentiment system is significantly better than the Naive Baye's sentiment system across all three categories. From a performance perspective, this shows that the proposed ME system is competitive.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2678164, IEEE/ACM Transactions on Audio, Speech, and Language Processing

B. Audio Sentiment Results

In the first audio experiment, we try to determine the best keyword list for audio based sentiment detection. By using all keyword lists extracted using the method described in Section V-A, we evaluate sentiment detection accuracy for both YouTube and UT-Opinion datasets against each keyword list. Figure 6 shows sentiment accuracy *vs.* keyword list size. It shows the impact of choosing larger sized keyword sets in order to improve the overall system accuracy. The intention is to show the saturating behavior of accuracy as we choose more and more keywords. It is observed that accuracy for both YouTube and UT-Opinion first increases as we increase the keyword list size from 200 to 32K. After 32K, the accuracy seems to level off before eventually decreasing. Since the smallest keyword lists have the most effective sentiment bearing features, the accuracy initially rises quickly.

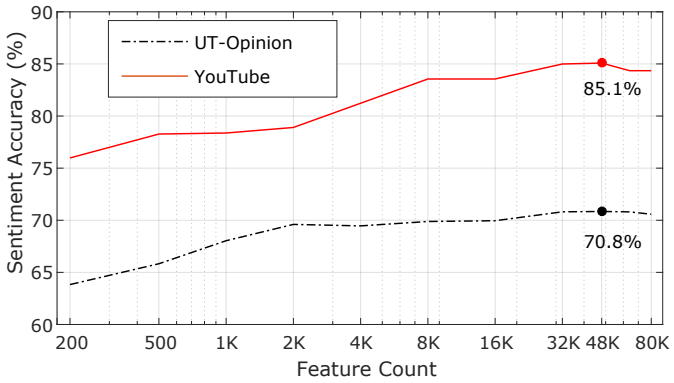

Fig. 6. Accuracy distribution curve for proposed hybrid ME-KWS audio sentiment system. This gives information regarding the most optimal keyword list to be used.

As the keyword size grows, the new keywords are relatively less effective, and therefore accuracy does not rise as rapidly. In KWS systems, it is commonly understood that increasing the size of the keyword list can potentially increase the number of false alarms (*i.e.*, the system falsely detects a keyword that is not actually present in the audio). In other words, bigger lists can be less accurate leading to high false alarms. From the trend in Figure 6, it appears that incorrectly detected features (due to errors in KWS) in the biggest keyword lists overwhelm the system resulting in a loss of accuracy. It is possible that a high number of false alarms contribute towards this outcome. Since the 48K keyword set gives the best performance, we choose this keyword set for all remaining experiments.

TABLE IV
RESULTS FOR THE PROPOSED HYBRID ME-KWS AND ASR TRANSCRIPT+ME SYSTEM.

Dataset	**Method**	
	ASR Transcripts with Max Entropy	**Proposed Hybrid ME+KWS (Total Error Reduction)**
YouTube	18.41%	**14.9% (19%)**
UT-Opinion	31.8%	**29.2% (8.2%)**

Table IV shows the EER of the baseline ASR transcript based sentiment system and the proposed Hybrid ME-KWS based sentiment detection systems. It can be observed that the proposed method outperforms the baseline system on both YouTube and UT-Opinion corpora.

The proposed system yields a +4% absolute improvement on the YouTube dataset (19% relative reduction in error) and a +2.6% absolute improvement on UT-Opinion dataset (8.2% relative reduction in error). Additionally, the performance of both ASR and KWS systems are better on the YouTube dataset when compared to UT-Opinion. The results show that the proposed KWS method for sentiment detection outperforms the baseline ASR method for both corpora. The results also show that sentiment detection task on UT-Opinion is more difficult, perhaps owing to wider variations in accents, noise, and subtly in sentiment expression. Another main reason is that UT-Opinion audio files are shorter in duration and captured in field interview scenarios. Since YouTube files are longer, there is more evidence for consistent sentiment detection, making the task relatively easier. In YouTube, in general, people express stronger opinion than the UT-opinion corpus (we note that the topics in UT-opinion are generally less controversial than many examples in YouTube), which also makes the problem relatively easier.

TABLE V
PERFORMANCE ACROSS VARIOUS PARAMETERS IN UT-OPINION DATABASE.

Variable	**Type**	**Person Count**	**Sentiment Detection EER**	
			ASR Transcript	**Proposed Hybrid ME-KWS**
Accent	Native	40	25.8%	23.5%
	Non Native	23	38.1%	34.9%
Environment	Office	41	20.6%	19.1%
	Hallway + Library + Gym	12	43.0%	39.3%

Table V gives a detailed distribution of performance across various parameters of the developed UT-Opinion database. The performance is discussed in terms of accent and environment. It can be observed that the performance is improved in both the cases by using the proposed method. There is a very good improvement in detecting non-native speaker's sentiment using ME-KWS system. Hallway and Gym are very noisy places. The performance in such environments have also increased by the using the proposed method. Human inter-rater agreement being relatively low for UT-Opinion as shown in Section VII-B is further proof that the sentiment detection for this data is a challenge.

But, when compared to YouTube audio database and text based systems the performance is almost less by 15%. Performance on text corporas are high because the data is rich in opinion. In text based reviews, a case where a reviewer doesn't have a strong opinion are seldom less. Reviews with strong sentiment based words are easier to classify. Similarly, in YouTube database the opinions of the expressed by particularly strong and detailed. Hence the performance of the YouTube

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2017.2678164, IEEE/ACM Transactions on Audio, Speech, and Language Processing

is relatively comparable to text based systems given some issues introduced by either speech recognition or speech signal related issues. But in UT-opinion the question are more natural and things that people come across in their everyday life.

IX. Conclusion

In this study, a new method for recognizing sentiment in audio has been proposed. The analysis shows that overall sentiment (both in audio and text) is governed by few sentiment bearing terms. In order to exploit this fact, a new method that uses Keyword Spotting (KWS) to search for sentiment bearing terms in audio has been proposed. By focussing on the terms that impact decision and ignoring non-sentiment bearing words/phrases, the overall system is more immune to speech recognition errors. Additionally, a new method to create the sentiment bearing keyword list for KWS has also been proposed. The method uses an iterative methodology to automatically extract sentiment bearing keywords from text. Using this method, we are able to build more practical systems that utilize equal to or less than 48K keywords. Additionally, a new method for sentiment scoring that combines keyword spotting likelihood (or confidence) into Maximum Entropy likelihood computation has also been proposed. Furthermore, a new corpus for audio sentiment evaluation has been collected and presented in this study. The new corpus is called UT-Opinion and to the best of our knowledge, is one of its kind for audio based sentiment detection. Finally, we have presented the evaluation of the proposed system on YouTube and UT-Opinion corpora. The new method has been compared to a baseline system that uses raw transcripts from ASR and feeds it to text based sentiment classifier. Our experimental results show that the new method outperforms the baseline system by reducing the error rate by 19% relative in YouTube, and 8% relative in UT-Opinion. While the new method improves upon audio based sentiment detection, there is room for further improvement. For example, addressing the traditional robustness problems of ASR (accent, noise, *etc.*) can have significant impact of performance. Another areas of work could focus on using pure speech features to augment lexical information drawn for speech recognition to do speech sentiment detection.

References

[1] S. Johnson, "Internet changes everything: Revolutionizing public participation and access to government information through the Internet", Administrative Law Review, Vol. 50, No. 2 (Spring 1998) , pp. 277-337

[2] D. Chrysanthos. "Strategic manipulation of internet opinion forums: Implications for consumers and firms." Management Science 52.10 (2006): 1577-1593.

[3] M. Wollmer, et al. "Youtube movie reviews: Sentiment analysis in an audio-visual context." Intelligent Systems, IEEE (2013): pages 46-53.

[4] J. Naughton, "The internet: is it changing the way we think?", The Gaurdian, Saturday 14 August 2010

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

[6] L. Barbosa, and J. Feng, "Robust sentiment detection on twitter from biased and noisy data.", in Proceedings of the International Conference on Computational Linguistics (COLING-2010). 2010.

[7] E. Cambria, N. Howard, Y. Xia, and T. S. Chua, "Computational Intelligence for Big Social Data Analysis", IEEE Computational Intelligence Magazine, 11(3), 8-9, 2016.

[8] E. Cambria, B. Schuller, Y. Xia, and B. White, "New avenues in knowledge bases for natural language processing", Knowledge-Based Systems, 108(C), 1-4, 2016.

[9] M. Bautin, L. Vijayarenu, and S. Skiena. "International sentiment analysis for news and blogs.", in Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM-2008). 2008.

[10] I. Becker and V. Aharonson. "Last but definitely not least: on the role of the last sentence in automatic polarity-classification.", in Proceedings of the ACL 2010 Conference Short Papers. 2010.

[11] Z. Zhai, B. Liu, H. Xu and P. Jia. "Clustering Product Features for Opinion Mining." Proceedings of Fourth ACM International Conference on Web Search and Data Mining (WSDM-2011), Feb. 9-12, 2011, Hong Kong, China.

[12] L. Kaushik, A. Sangwan, and J. H. L. Hansen. "Sentiment extraction from natural audio streams." In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 8485-8489. IEEE, 2013.

[13] L. Kaushik, A. Sangwan, and J. H. L. Hansen. "Automatic sentiment extraction from YouTube videos." In Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on, pp. 239-244. IEEE, 2013.

[14] 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

[15] E. Kouloumpis, T. Wilson, and J. Moore. "Twitter sentiment analysis: The good the bad and the omg!." ICWSM 11 (2011): 538-541.

[16] A. Bifet, and E. Frank. "Sentiment knowledge discovery in twitter streaming data." In Discovery Science, pp. 1-15. Springer Berlin Heidelberg, 2010.

[17] H. Saif, Y. He, and H. Alani. "Semantic sentiment analysis of twitter." In The Semantic WebISWC 2012, pp. 508-524. Springer Berlin Heidelberg, 2012.

[18] A. Go, L. Huang, and R. Bhayani. "Twitter sentiment analysis." Entropy 17 (2009).

[19] A. Berger, D. Pietra, and D. Pietra, "A Maximum Entropy Approach To Natural Language Processing.", Computational Linguistics

[20] X. Fei, H. Wang, J. Zhu, "Sentiment word identification using the maximum entropy model," Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on , vol., no., pp.1,4, 21-23 Aug. 2010

[21] A. Ratnaparkhi, "A simple introduction to maximum entropy models for natural language processing.", IRCS Technical Reports Series (1997).

[22] C. Lin, and Y. He, "Joint sentiment/topic model for sentiment analysis.", 2009, Proceedings of the ACM International Conference on Information and Knowledge Management.

[23] B. Liu, "Sentiment Analysis and Opinion Mining.", 2012, Morgan & Claypool publishers

[24] Q. Mei, X. Ling, M. Wondra, H. Su and C. Zhai. "Topic sentiment mixture: modeling facets and opinions in weblogs." 2007, Proceedings of International Conference on World Wide Web.

[25] Y. Lu, and C. Zhai, "Opinion integration through semi-supervised topic modeling." Proceedings of International Conference on World Wide Web, 2008.

[26] S. Moghaddam, and E. Martin, "ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews.", In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 665-674. ACM, 2011.

[27] A. Mukherjee and B. Liu, "Mining contentions from discussions and debates.", In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12). ACM, New York, NY, USA, 841-849, 2012.

[28] M. Hu, and B. Liu. "Mining and summarizing customer reviews." In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177. ACM, 2004.

[29] W. Zhao, J. Jing, Y. Hongfei, and L. Xiaoming, "Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid.", In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 56-65. Association for Computational Linguistics, 2010.

[30] B. Pang, and L. Lillian, "Opinion mining and sentiment analysis." Foundations and trends in information retrieval 2.1-2 (2008): 1-135.

[31] B. Snyder, and R. Barzilay, "Multiple Aspect Ranking Using the Good Grief Algorithm." In HLT-NAACL, pp. 300-307. 2007.

[32] M. Joshi, and C. Penstein-Ros, "Generalizing dependency features for opinion mining." Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Association for Computational Linguistics, 2009.

[33] X. Ding, B. Liu, and P. S. Yu. "A holistic lexicon-based approach to opinion mining." In Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 231-240. ACM, 2008.

[34] L. Qian, Z. Gao, B. Liu, and Y. Zhang. "Automated rule selection for aspect extraction in opinion mining." In International Joint Conference on Artificial Intelligence (IJCAI). 2015.

[35] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts,", Proc. 42nd Ann. Meeting of the Assoc. for Computational Linguistics, Assoc. for Computational Linguistics, 2004, pp. 271278.

[36] E. Cambria and A. Hussain, "Sentic Computing: Techniques, Tools, and Applications", Springer, 2012.

[37] Witold Pedrycz, and Shyi-Ming Chen. "Sentiment Analysis and Ontology Engineering.", Springer, 2016

[38] V. Hatzivassiloglou, and K. R. McKeown. "Predicting the semantic orientation of adjectives." In Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics, pp. 174-181. Association for Computational Linguistics, 1997.

[39] B. Pang, L. Lee, and S. Vaithyanathan, Thumbs Up? Sentiment Classification Using Machine Learning Techniques, Proc. Ann. Conf. Empirical Methods in Natural Language Processing, Assoc. for Computational Linguistics, 2002, pp. 7986.

[40] X. Bai, "Predicting consumer sentiments from online text.", Decision Support Syst 2011;50:73242..

[41] A. Popescu, and E. Orena, "Extracting product features and opinions from reviews." In Natural language processing and text mining, pp. 9-28. Springer London, 2007.

[42] P. Turney, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, Proc. 40th Ann. Assoc. for Computational Linguistics, Assoc. for Computational Linguistics, 2002, pp. 417424.

[43] J. Kamps et al., Using WordNet to Measure Semantic Orientation of Adjectives, Proc. 4th Ann. Intl. Conf. IS-28-02-Cambria.indd 20 6/5/13 11:05 AM march/april 2013 www.computer.org/intelligent 21 Language Resources and Evaluation, European Language Resources Assoc., 2004, pp. 11151118

[44] R. Feldman, "Techniques and applications for sentiment analysis." Communications of the ACM 56.4 (2013): 82-89.

[45] J. Brooke, M. Tofiloski, and M. Taboada. "Cross-Linguistic Sentiment Analysis: From English to Spanish." RANLP. 2009.

[46] X. Wan, "Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis." Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2008.

[47] H. Yu, and V. Hatzivassiloglou. "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences." Proceedings of the 2003 conference on Empirical methods in natural language processing. Association for Computational Linguistics, 2003.

[48] E. Riloff , and J. Wiebe. "Learning extraction patterns for subjective expressions." Proceedings of the 2003 conference on Empirical methods in natural language processing. Association for Computational Linguistics, 2003.

[49] R. Socher, et al. "Recursive deep models for semantic compositionality over a sentiment treebank." Proceedings of the conference on empirical methods in natural language processing (EMNLP). Vol. 1631. 2013.

[50] T. Wilson, J. Wiebe, and P. Hoffmann. "Recognizing contextual polarity in phrase-level sentiment analysis." In Proceedings of the conference on human language technology and empirical methods in natural language processing, pp. 347-354. Association for Computational Linguistics, 2005.

[51] X. Ding, B. Liu, and L. Zhang. "Entity discovery and assignment for opinion mining applications." In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1125-1134. ACM, 2009.

[52] M. Ruiz, and P. Srinivasan. "Hierarchical neural networks for text categorization (poster abstract)." In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 281-282. ACM, 1999.

[53] L. Kaushik, A. Sangwan, and J. H. L. Hansen. "Automatic audio sentiment extraction using keyword spotting." In INTERSPEECH, pp. 2709-2713. 2015.

[54] N. H. Tou, G. Wei, L. Kok. "Feature selection, perceptron learning, and a usability case study for text categorization", ACM SIGIR conference; 1997

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

[56] http://www.cs.uic.edu/ liub/FBS/opinion-lexicon-English.rar

[57] J. McAuley, C. Targett, Q. Shi, and A. D. Hengel. "Image-based recommendations on styles and substitutes." In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 43-52. ACM, 2015.

[58] J. McAuley, R. Pandey, and J. Leskovec. "Inferring networks of substitutable and complementary products." In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794. ACM, 2015.

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

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

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

[62] K. Toutanova and C. D. Manning. 2000. "Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger" EMNLP/VLC-2000, pp. 63-70.

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

[64] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, The SEMAINE database: Annotated multi- modal records of emotionally coloured conversations between a person and a limited agent, IEEE Transactions on Affective Computing, 2011.

[65] www.youtube.com

[66] L. Morency, R. Mihalcea, and P. Doshi. "Towards multimodal sentiment analysis: Harvesting opinions from the web." Proceedings of the 13th international conference on multimodal interfaces. ACM, 2011.

[67] E. Cambria, B. Schuller, Y. Xia, and C. Havasi. "New avenues in opinion mining and sentiment analysis." IEEE Intelligent Systems 2 (2013), Pages 15-21.

[68] N. Jindal, B. Liu. "Identifying comparative sentences in text documents." In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 244-251. ACM, 2006.

[69] H. Cui, V. Mittal, and M. Datar. "Comparative experiments on sentiment classification for online product reviews." In AAAI, vol. 6, pp. 1265-1270. 2006.

[70] A. Sangwan, and J. H. L. Hansen. "Keyword recognition with phone confusion networks and phonological features based keyword threshold detection." In Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on, pp. 711-715. IEEE, 2010.

[71] J. Keshet, D. Grangier, and S. Bengio. "Discriminative keyword spotting." Speech Communication 51, no. 4 (2009): 317-329.

[72] D. Povey, et al. "The Kaldi speech recognition toolkit." (2011).

[73] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development.", In ICASSP, pp. 517520, 1992.

[74] C. C. David, D. Miller, and K. Walker. The fisher corpus: a resource for the next generations of speech-to-text. In in Proceedings 4th International Conference on Language Resources and Evaluation, pages 6971, 2004.

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