Automatic Audio Sentiment Extraction Using Keyword Spotting

Conference Paper · September 2015

DOI: 10.21437/Interspeech.2015-571

Lakshmish Kaushik
University of Texas at Dallas
28 PUBLICATIONS **335** CITATIONS

SEE PROFILE

Abhijeet Sangwan
Speetra
89 PUBLICATIONS **954** CITATIONS

SEE PROFILE

John H. L. Hansen
University of Texas at Dallas
660 PUBLICATIONS **13,761** CITATIONS

SEE PROFILE

Automatic Audio Sentiment Extraction Using Keyword Spotting

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

Most existing methods for audio sentiment analysis use automatic speech recognition to convert speech to text, and feed the textual input to text-based sentiment classifiers. This study shows that such methods may not be optimal, and proposes an alternate architecture where a single keyword spotting system (KWS) is developed for sentiment detection. In the new architecture, the text-based sentiment classifier is utilized to automatically determine the most powerful sentiment-bearing terms, which is then used as the term list for KWS. In order to obtain a compact yet powerful term list, a new method is proposed to reduce text-based sentiment classifier model complexity while maintaining good classification accuracy. Finally, the term list information is utilized to build a more focused language model for the speech recognition system. The result is a single integrated solution which is focused on vocabulary that directly impacts classification. The proposed solution is evaluated on videos from YouTube.com and UT-Opinion corpus (which contains naturalistic opinionated audio collected in real-world conditions). Our experimental results show that the KWS based system significantly outperforms the traditional architecture in difficult practical tasks.

Index Terms:Audio sentiment detection, Reviews, Maximum Entropy, KWS, KALDI, NLP, ASR, YouTube, UT-Dallas Opinion Audio Archive

1. Introduction

In this study, a new system for sentiment detection in audio is presented. While automatic sentiment detection using text is a mature area of research, and significant research has been done on product reviews [1, 2, 3, 4, 5], audio sentiment detection remains relatively under explored. Given the explosive increase of online videos on product reviews, un-boxing, politics, sports, culture, *etc.* on websites such as YouTube.com, automatic audio sentiment detection technology would undoubtedly be useful is collecting and summarizing information for users.

It is useful to note that audio sentiment detection concerns with detection of opinion (positive *vs.* negative), and is different from speech emotion recognition. The popular architecture employed by researchers studying audio based sentiment detection is a tandem system that utilizes automatic speech recognition (ASR) technology to convert speech into text, following by conventional text-based sentiment detection systems [16, 8, 9]. In this manner, the text processing system searches for sentiment bearing features (words, phrases, *etc.*) in the output of the speech recognizer.

Accurate sentiment detection 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 ordered a pepperoni pizza last night and it was wonderful”, only 1 out of 11 words conveys sentiment. While this may not be true for every comment, 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 high Word Error Rates (WERs), and this is precisely what we have observed in previous studies [9]. In other words, sentiment detection accuracy depends on being able to reliably detect a very focused vocabulary in the spoken comments. Therefore, Keyword Spotting (KWS) technology seems to be better suited for sentiment detection, as opposed to full-transcript ASR.

In order to build an effective KWS system, we need an compact yet effective keyword list. The textual features extracted by most text-based sentiment classification system is a good starting point to generate a keyword list. However, the learning paradigm for most of these systems tends to be greedy and generates a very large number of features. To mitigate this problem, we propose an iterative technique that can reduce the feature size (and consequently model complexity) without significantly sacrificing performance accuracy. Additionally, we incorporate the term list in the speech recognition language model to assist in better KWS by ensuring none of the vital terms are OOV (Out of Vocabulary). The mentioned innovations deliver a single integrated system.

In this study, the proposed system is evaluated on two corpora: UT-Opinion and videos from youtube.com. UT-Opinion is a new corpus that we have specifically collected for the purpose of audio sentiment detection. The new corpus contains interview style data, where subjects give their opinion of various topics in natural settings. Our experimental results show that the proposed KWS framework significantly outperforms the conventional ASR approach on both tasks.

2. Proposed System

Figure 1 shows the proposed KWS-based and traditional ASR-based approaches for sentiment analysis. In both systems, ASR language model is prepared offline and is trained (or adapted) for the domain. Next, speech recognition is applied to the audio data. In traditional sentiment detection systems, 1-best transcripts are obtained from speech recognition and fed to the text-based sentiment detection unit which provides final classification. In the new system, we propose to extract lattices from the ASR stage, and use KWS to search for sentiment bearing terms alone. The term list for KWS is generated offline from a large text collection (of reviews, opinions, *etc.*) using a new iterative algorithm described below.

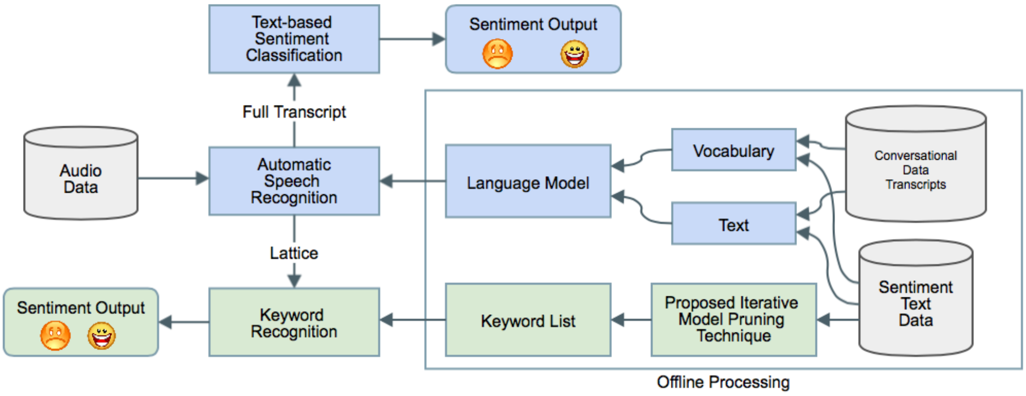

Figure 1: Block diagram for the proposed audio based sentiment system. The language model for the speech recognizer is built offline by using a mixture of sentiment text data and conversational telephony transcripts. Sentiment text data is also used to generate the keyword term list by applying the proposed iterative pruning method. The proposed sentiment detection system uses the term list to search for sentiment bearing terms in the audio.

2.1. Keyword Generation

2.1.1. Text Based Sentiment Classifier

Our text based sentiment classifier is based on Maximum Entropy (ME) modeling technique. The ME features are extracted from the text by selecting textual features corresponding to adjectives, adjective-noun pairs, noun-clusters, *etc.* using Part-of-Speech tagging (POS). More details can be found in [8, 9].

The ME model is trained to predict sentiment given textual features extracted from the comment. 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:

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

Now, the ME technique can predict the rating of the review y_j from features x_k by using:

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

where, $Z_\lambda(x)$ is a normalizing term, andλ_{ij} are weights assigned to thef_i (and are learned during training).

In this study, a number of text sources such as Amazon, Yelp, TripAdvisor, Pros & Cons data, Scale data, Comparative Data, etc., are used for generating training data [9]. Altogether, the dataset contains 6m reviews, and after training, the model contains close to a million unique text-based features. While each text based feature could be a potential keyword, the list as such is very large and not suitable for KWS. Therefore, we develop a pruning method that can dramatically reduce the features while maintaining accuracy. In what follows, we explain this new method.

2.1.2. Iterative Model Pruning Method

Figure 2 shows the proposed iterative threshold based pruning approach which is used to generate the necessary keyword list. As shown in the figure, we first extract POS-based textual features that are potentially sentiment-bearing. We focus on extracting textual features that contain adjectives, nouns, noun-clusters, adjective-noun pairs, *etc.* In this study, we used the Stanford POS tagger [6, 7]. These steps provide the initial feature set of the iterative method. As mentioned previously, this feature list tends to be large and needs to be pruned to capture the most effective terms (words, phrases, *etc.*).

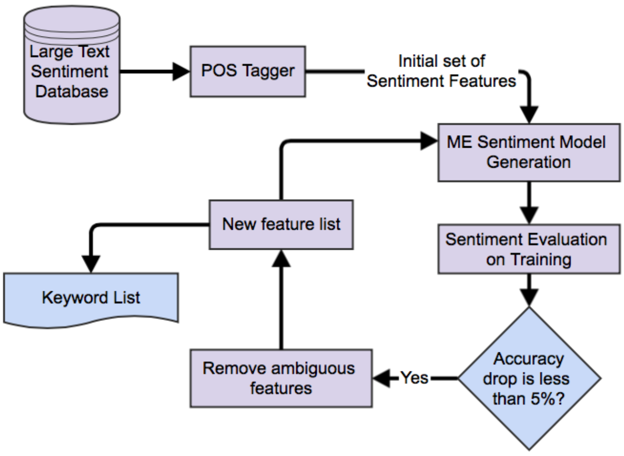
Large Text Sentiment Database
POS Tagger
Initial set of Sentiment Features
ME Sentiment Model Generation
New feature list
Sentiment Evaluation on Training
Keyword List
Remove ambiguous features
Yes
Accuracy drop is less than 5%?

Figure 2: Iterative threshold pruning to generate Sentiment Models and Keyword list for the proposed Keyword Spotting based Sentiment Detection System

Next, the feature list is used to train the ME based sentiment model. Once the model is trained, we can probe the model with individual features to compute it's probability of positive (or negative) sentiment. In other words, the ME model learns conditional probabilities of sentiment given feature, and we can exploit this information to remove ambiguous features. For example, if a feature probability of positive sentiment is 0.5, then it is not discriminative in selecting a sentiment. We define a probability range to identify ambiguous features, *i.e.*, a feature is ambiguous if the probability of positive sentiment lies between α_{lower} and α_{upper}. In addition to the mentioned constraint, we also enforce that any feature that occurs less than N times in the training set is ambiguous. The combination of mentioned constraints ensure that features that are infrequent and/or non-discriminative are rejected. The remaining features are retained for the next iteration. In this study, we used $\alpha_{lower} = 0.45$ and $\alpha_{upper} = 0.55$.

In the next iteration, the features from the previous iteration are used to train a new ME sentiment model. Ambiguous feature identification is repeated with one exception, *i.e.*, we cast a wider net for ambiguous features by decrementing α_{lower} and incrementing α_{upper} by 0.05, respectively. This process is repeated in subsequent iterations, where α_{lower} and α_{upper} values are successively decremented and incremented by 0.05, respectively. This continues till α_{lower} and α_{upper} become equal to 0.3 and 0.7, respectively. At this point, we continue subsequent iterations without decreasing and increasing the values of α_{lower} and α_{upper}.

With successive iterations, the total number of features fall dramatically with some fall in classification accuracy on the training dataset. For this study, we continue the iterations till our classification accuracy falls to just below 5% of the baseline accuracy or the difference in the total number of features from the present and previous iteration is less than 100 sentiment features. At this point, the remaining features are used as the term list for the KWS system. The thresholds used (5% of baseline or difference less than 100) is chosen keeping in mind not to compromise on the the text sentiment models accuracy and at the same time help to develop a concise and efficient KWS sentiment keyword list.

2.2. Automatic Speech Recognition (ASR)

A Kaldi based ASR system was used for evaluation [12]. This baseline system is a conventional ASR continuous speech recognizer. The parameters and data in acoustic model and language models are described below.

2.2.1. Acoustic and Language 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). The acoustic model for evaluation was trained using mix-style approach, where acoustic data from multiple corpora were used. Part of the training data was corrupted with additive noise at various SNRs (signal to noise ratio) to recreate the various condition we expect to see in the evaluation corpora used in this research work.

Standard triphone based Hidden Markov Models (HMMs) are used for this study. In the feature space, standard MFCC (Mel Frequency Cepstral Coefficients) features with delta and delta-delta coefficients are adopted. 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. In the next step, feature transforms like 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.

The speech recognition vocabulary contained 20,000 most frequently occurring words (determined from conversational telephony transcripts). Additionally, vocabulary from the keyword list generated using the proposed iterative method was also included in the speech recognizer.

The language model was trained on data from two sources, namely, CTS (conversational telephone speech) data from Switchboard and Fisher corpora, and reviews data (which is a collection of various review/opinion datasets such as Amazon, TripAdvisor, *etc.*) [9]. This allowed us to model the contextual dependencies of all the target keywords.

A trigram language model was trained using the following data sources: (i) Amazon Product Reviews [10] (5.8million reviews which is approximately 2billion words), (i) Switchboard, (ii) Fisher (iii) UW191 [11] (191M words collected from the web by the University of Washington), and (iv) other sentiment datasets mentioned in Sec. 2.1.

During decoding, we executed two rounds of fMLLR transform estimation, before using the second pass fMLLR transform for rescoring the decoded lattices. For the conventional ASR based sentiment system, one-best transcripts from the recognizer were extracted. For the proposed KWS system, word lattices and corresponding phone lattices were extracted as ASR output.

2.3. Keyword Spotting

Using the word lattice generated by the ASR in the previous step, a Finite State Transducer (FST) based method was used to search the word lattices for keywords [14, 15]. In parallel, the word lattices were converted into phone lattices, and the PCN-KWS (phone confusion network keyword spotting) method was employed to search for keywords [13]. Subsequently, the search results from the two methods were combined (by simple likelihood combination) to yield the final keyword result list.

3. Evaluation Corpora

3.1. YouTube Corpus

A set of 80 videos were collected from YouTube.com as a part of natural sentiment database collection. These videos cover a wide range of topics including product reviews, movies, social issues and political opinions. The audio quality, recording equipment, channel characteristics, and accents/dialects vary across videos. More details of the database are discussed in Table 1 of results section. These videos can be accessed via a YouTube playlist: http://bit.ly/YAgoYU. More details about the database can be found in [9].

3.2. UT-Opinion Corpus

While YouTube is good resource for sentiment videos, it has some disadvantages such as: (i) it is hard to consistently find multiple opinionated videos for a single speaker (with reasonable topic diversity), (ii) it is harder to balance factors such as age, gender, nativeness *etc.*, and (iii) neutral sentiment (or "don't care") is hard to find in a meaningful way as people only post videos when they have a strong opinion and are motivated enough to express them. Therefore, a more controlled collection which retains the naturalness of YouTube but addresses some of the above concerns is beneficial. We collected UT-Opinion corpus for this reason.

In UT-Opinion, each subject is interviewed where they are asked to respond to 10 questions. The questions have been designed to illicit opinions. After sharing their spoken comments, the participants are asked to rate their sentiment for every question on a five point scale, *i.e.*, strongly positive, positive, neutral, negative, and strongly negative. Subjects are interviewed at various locations on the University of Texas at Dallas (UTD) campus including classrooms, hallway, office rooms, library, gym and street. The subjects included students, staff and faculty members of both gender. Both native and non-native speakers were included in the collection. Altogether, data from 120 subjects has been collected resulting in 1200 evaluation audio files.

4. Results and Analysis

In the first experiment, we study the proposed iterative feature reduction strategy. Figure 3 shows the reduction is feature set with every iteration, and the corresponding reduction in accuracy (on the train set itself). It can be seen in the figure that a huge reduction in feature size was obtained within 6 iterations (we had 687K after first iteration features and 12,500 features after the last iteration). The corresponding classification accuracy drops by 5.33%.

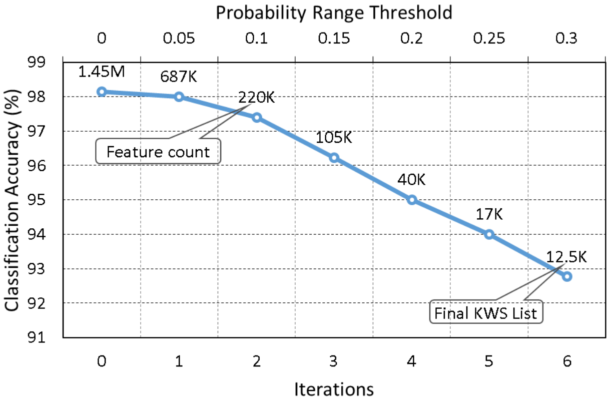

Figure 3: In the Iterative threshold pruning technique, the number of features are dramatically reduced with every iteration with a small decrease in training classification accuracy. Numbers at each node represent the #of features at that iteration. 12,500 features remain at 6^{th} iteration and are used as the keyword term list.

In the next experiment, we evaluated both the ASR and KWS based sentiment detection methods on YouTube and UT-Opinion datasets. Table 2 shows the performance accuracy of sentiment detection using both methods and databases. From the table, we can observe that the proposed KWS system outperforms the traditional ASR-based method. For YouTube.com videos and UT-Opinion corpus, the performance accuracy of the KWS system is 5.7% (absolute) and 12.3% (absolute) better than the traditional ASR-based system, respectively. Additionally, it is also observed that the overall performance for UT-Opinion is lower than YouTube.com videos.

Corpora information for YouTube and UT-Opinion are given in Table 1 below. It can be seen that both corpora contain sufficient amount of data and good speaker diversity. UT-Opinion is more challenging than YouTube in terms of accent diversity.

Table 1: The duration, sentiment, gender, environment, and accent information for both corpora are shown below.

	UT-Opinion Database	YouTube Sentiment Database
Total Duration	12.68Hours	7.5Hours
Total Speakers	120	85
Gender	Males (68), Females (52)	Males (50), Females (30)
Environment	Office, Hallway, Library, Gym	Unknown locations, Env with echo, Varied recording device setups
Speaker Accents	American, Indian, Farsi, Arabic, Chinese, Korean, Hispanic, Italian, French	America, Hispanic, British
Sentiment Count (%)	Positive: 54% Negative: 46%	Positive: 51.77% Negative: 48.23%

In general, YouTube videos are made by speakers who are motivated and generally more expressive about their opinion. The sentiment expressed in these videos is typically stronger than those expressed in UT-Opinion.

Table 2: Best Sentiment detection accuracy for traditional ASR and proposed KWS method.

Corpus	Traditional ASR Method	Proposed KWS Method
YouTube	85.3%	91.1%
UT-Opinion	56.4%	68.7%

Additionally, a number of subjects in UT-Opinion either did not have a strong opinion on a number of topics, or did not have an opinion at all. Hence, as such the UT-Opinion dataset represents a harder task. However, we feel that both YouTube and UT-Opinion datasets are practical as they cover the full range of expression that one can expect to see in practice.

Furthermore, the Detection Error Tradeoff (DET) curves for the UT-Opinion corpus using the proposed KWS system is shown in Figure 4. From the DET curves, it is seen that the proposed system outperforms the traditional ASR-method at all operating points.

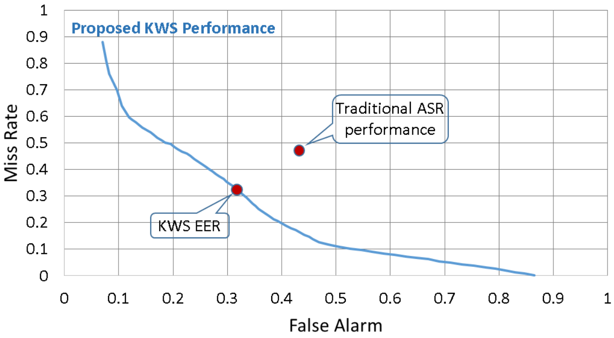

Figure 4: DET (Detection Error Tradeoff) curve showing performance of the proposed KWS based sentiment detection system for UT-Opinion corpus.

We can observe from Figure 4 that EER point for KWS based sentiemnt system is 0.32. The gains due to the proposed KWS system are remarkable given that the traditional ASR system performance is close to random (i.e, a system that guesses positive and negative labels).

5. Conclusion

A new method for audio sentiment detection based on KWS has been presented. The new method exploits the fact that the lexical evidence for sentiment in spoken comments is sparse and depends on a relatively smaller focused vocabulary. Unlike traditional audio sentiment detection solution, which combines full-transcript ASR with text-based sentiment processing system sentiment, the proposed solution offers a single integrated solution. It is possible that the proposed KWS architecture may be suitable for other high-level semantic classification tasks as well. The new method was evaluated on practical data from YouTube.com and UT-Opinion corpus, and was shown to outperform the traditional ASR approach by 12% absolute increase in classification accuracy.

6. Acknowledgement

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

7. References

[1] C. Lin and Y. He, “Joint sentiment/topic model for sentiment analysis,” in International Conference on Information and Knowledge Management, pp. 375-384, 2009.

[2] B. Liu, “Sentiment Analysis and Opinion Mining,” Morgan and Claypool publishers, 2012.

[3] Y. Lu and C. Zhai, ”Opinion integration through semi-supervised topic modeling”, in International Conference on World Wide Web, pp. 121-130, 2008.

[4] S. Moghaddam and M. Ester, “ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews”, in SIGIR Conference on Research and Development in Information Retrieval, pp. 665-674, 2011.

[5] M. Arjun and B. Liu, “Mining contentions from discussions and debates”, in SIGKDD international conference on Knowledge discovery and data mining (KDD ’12), pp. 841-849, 2012.

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

[7] K. Toutanova, D. Klein, C. D. Manning and Y. Singer, “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network,” in HLT-NAACL 2003, pp. 252-259, 2003.

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

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

[10] Z. Zhongwu, B. Liu, H. Xu and P. Jia, ”Clustering product features for opinion mining.” in ACM international conference on Web search and data mining, pp. 347-354, 2011.

[11] "http://ssli.ee.washington.edu/ssli/projects/ears/WebData/web_data_collection.html''

[12] P. Daniel, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann et al. ”The Kaldi speech recognition toolkit.”, 2011.

[13] A. Sangwan and J.H.L. Hansen, Keyword recognition with phone confusion networks and phonological features based keyword threshold detection, in Asilomar Conference on Signals Systems and Computers (ASILOMAR), pp. 711715, Nov. 2010.

[14] M. Arindam, J. Hout, Y. Tam, V. Mitra, Y. Lei, J. Zheng, D. Vergyri et al. ”Strategies for high accuracy keyword detection in noisy channels.” In INTERSPEECH, pp. 15-19. 2013.

[15] M. Akbacak, L. Burget, W. Wang and J. V. Hout, “Rich system combination for keyword spotting in noisy and acoustically heterogeneous audio streams”, in ICASSP 2013, pp. 8267-8271, 2013

[16] S. Ezzat, N. Gayar and M.M. Ghanem, ”Sentiment Analysis of Call Centre Audio Conversations using Text Classification,” in International Journal of Computer Information Systems and Industrial Management Applications, vol. 4, pp. 619-627, 2012.