1-1-2013

Temporal sentiment detection for user generated video product reviews

M S. Barakat
University of Wollongong, mb452@uowmail.edu.au

C H. Ritz
University of Wollongong, critz@uow.edu.au

D A. Stirling
University of Wollongong, stirling@uow.edu.au

Recommended Citation

Barakat, M S.; Ritz, C H.; and Stirling, D A., "Temporal sentiment detection for user generated video product reviews" (2013). *Faculty of Engineering and Information Sciences - Papers: Part A*. 1949.
https://ro.uow.edu.au/eispapers/1949

Temporal sentiment detection for user generated video product reviews

Abstract

User generated video product reviews in social media is gaining popularity every day due to its creditability and the broad evaluation context provided by it. Extracting sentiment automatically from such videos will help the consumers making decisions and producers improving their products. This paper investigates the feasibility of sentiment detection temporally from those videos by analyzing the transcription generated by a speech recognition system which was not investigated before. Another two main contribution for this paper is introducing a solution to the problem of fixed threshold estimation for the Naive Bayesian classifier output probabilities and irrelative text filtering for improving the sentiment classifcation. Various experiments indicated the proposed system can achieve an F-score of 0.66 which is promising knowing that the sentiment classifier offers an F-score of 0.78 provided that the input text is error free.

Keywords

temporal, user, product, detection, reviews, generated, sentiment, video

Disciplines

Engineering | Science and Technology Studies

Publication Details

M. S. Barakat, C. H. Ritz & D. A. Stirling, "Temporal sentiment detection for user generated video product reviews," in 13th International Symposium on Communications and Information Technologies: Communication and Information Technology for New Life Style Beyond the Cloud, ISCIT 2013, 2013, pp. 580-584.

temporal sentiment detection for user generated video product reviews

M. S. Barakat, C. H. Ritz, D. A. Stirling

ICT Research Institute / School of Electrical, Computer and Telecommunication Engineering,
University of Wollongong,
NSW, Australia.

mb452@uowmail.edu.au *critz@uow.edu.au* *stirling@uow.edu.au*

***Abstract*—User generated video product reviews in social media gaining popularity every day due to its creditability and the broad evaluation context provided by it. Extracting sentiment automatically from such videos will help the consumers making decisions and producers improving their products. This paper investigates the feasibility of sentiment detection temporally from those videos by analyzing the transcription generated by a speech recognition system which was not investigated before. Another two main contribution for this paper is introducing a solution to the problem of fixed threshold estimation for the Naïve Bayesian classifier output probabilities and irrelative text filtering for improving the sentiment classification. Various experiments indicated the proposed system can achieve an F-score of 0.66 which is promising knowing that the sentiment classifier offers an F-score of 0.78 provided that the input text is error free.**

Keywords— Users Video Blogs, Social Media, Automatic Speech Recognition (ASR), Sentiment Classification.

I. INTRODUCTION

Social networks like YouTube, Facebook and Twitter that allow any person to generate and share opinions, and discuss and comment on certain issues or topics, have become one of the most popular platforms in the web world [1]. Recent studies declared that YouTube has become the most popular website [2]. This popularity and the freedom of uploading information and opinions about any topic including products made social media a very powerful marketing tool that is being actively used by governments, major organizations and producers [1, 3, 4].

Many studies reported that prior to buying a product, consumers tend to spend more time searching the internet for reviews than with the retailers directly [4, 5]. Well known companies have started collecting and analysing the huge database of opinions available on social networks to improve their products [6]. Hence, there is a real need for automatic sentiment analysis by organizations and individuals [1, 3, 6, 7].

While sentiment analysis has become a very active research area in textual blogs and reviews [1, 3, 6-8], less research has been done into sentiment analysis of video blogs [9-11], which are videos recorded and published by ordinary users expressing their feelings and opinions about products, events and many other issues [9, 12]. Video blogs have recently become the most popular types of blog or reviews [2, 9] when compared to text-based reviews. One reason is increased perceived credibility due to the visibility of the source of the review (i.e. the blogger), especially if they appear as an ordinary unbiased end-user [4, 5]. Secondly, a study appraised that video product reviews, provide broader context to the evaluation that goes beyond the basic description, functionality and cost by providing a real demonstration [10]. User blogs are also attractive to companies, as they can quickly gain feedback on their products.

Hence, this paper focuses on temporal sentiment analysis of the spoken content of video product reviews to determine at which points in time the blogger is speaking positively or negatively. The limited previous work on product review videos focuses on classifying the whole video clips like in [11] while to the authors knowledge extent there was no previous investigation on the temporal classification. Temporal analysis is important as it can be used for efficiently searching for positive and negative specific opinions and aspects without watching the entire video. This will help the producer gaining not only an overall feedback about the products but knowing which parts and features in the products the reviewers liked or hated. Further, text-based user comments on the video are not linked to specific time-points and are often out of context, being about the video itself, such as production or quality, rather than the main topic (in this case the product) which will make them not very useful [2].

Focusing on product reviews, the proposed approach is based on analysing the video text transcription produced from an automatic speech recognition system using a text-based sentiment classifier. Included is an investigation into the effect of the speech recognition error on the sentiment detection accuracy as well as post-processing methods applied to the outputs of the sentiment classifier to improve the performance. Another important contribution in this paper is the clustering of sentiment probabilities, to solve the problem of choosing thresholds indicating positive, negative or neutral sentiment. This paper also suggests a text-filtering stage, to remove sentences that are irrelevant to the product and shows that this significantly improves the sentiment detection results.

Section II of this paper will give a summary of the related work. Section II will provide a description of the proposed idea and system. Section IV describes a set of experiments and

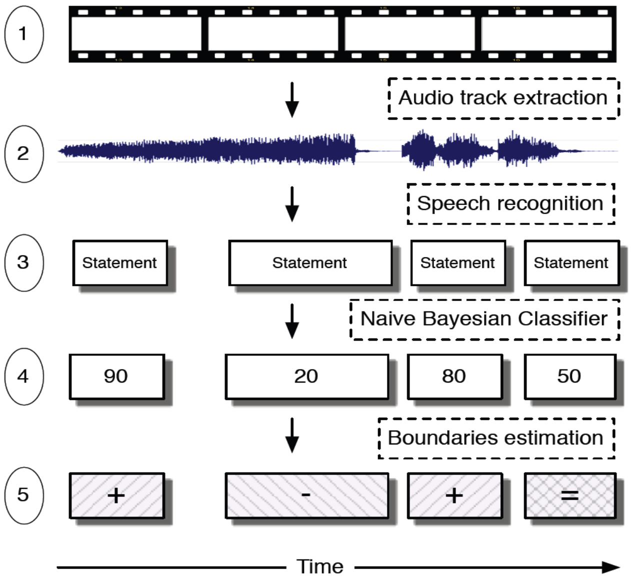

Fig. 1: The proposed temporal sentiment classification system

results evaluating the system, while Section V provides conclusions.

II. Related Work in Sentiment Analysis

Sentiment analysis or opinion mining is the computational study of people's opinions toward entities and their attributes [7]. The two main approaches to sentiment analysis are lexical or dictionary approaches and machine learning approaches [3, 6-8].

The lexical approach uses linguistic rules and predefined opinion word lists or dictionaries to classify text sentiment [7, 8]. Research has focused in ways of automatically generating and expanding these dictionaries and word lists [7, 8, 13, 14]. A known problem is word context, for example, the word "unpredictable" might indicate a negative opinion when it is used to describe a car and a positive one when describing a movie plot [7]. Further, the language in blogs usually informal and colloquial, which makes the use of lexicon based approaches very difficult [2, 3], requiring complicated natural language processing [7].

To overcome these limitations, building and training machine learning models with training text examples [3, 7, 8] has been proposed. The Naïve Bayesian classifier and support vector machine (SVM) are the most common approaches [7, 8], with the former approach providing the best results in many cases [8, 15, 16] and outperforming a lexical approach, even after applying more sophisticated language models [8]. A drawback of machine learning approaches is the need for large amounts of training data, but this is becoming more available now, especially in the form of product reviews[7, 17]. There were attempts to combine both approaches (lexical and machine learning), however the major improvement was reducing the amount of training data required for the classifier, even when using more complex language models [6, 8]. Hence, the Naïve Bayesian classifier was chosen for the system proposed in this paper.

III. The Proposed System

As shown in Figure 1, the audio signal is extracted from the video and introduced to the automatic speech recognition system (ASR). The resulting text statements (sentences) are entered as an input to the sentiment classifier. The classifier outputs the probability of each statement being positive or negative. The final step is processing these probabilities to determine class's boundaries and produce a final decision.
The ASR system selected for this system is the YouTube Automatic Transcriber which uses the same ASR in the popular commercial system Google Voice [18]. A recent comparison showed that the YouTube transcriber offers superior performance to an alternative speech recognizer (Pocketsphinx [19, 20]) when analyzing the spoken content of YouTube videos in [21].

Applying ASR to analyze videos was not previously successful due to the high word error rate [22], especially when the speech is non-read (spontaneous) speech [23, 24], such as the spoken content of user generated videos. However, since not every word is very important in sentiment analysis, it is believed that the error rate effect will be minimal.

The sentiment Naïve Bayesian classifier depends on calculating the probability of a class given a set of features based on the Bayes theorem and can be expressed as:

$$P(C|F) = \frac{P(F|C)P(C)}{P(F)} \qquad (1)$$

Where *F*= *{f1,...fn}*is the set of features (tokenized words of the input text) and C ={positive, negative}are the two classes [16, 25]. The Naïve Bayesian classifier used in this system [25] was trained using the Amazon product review database, using more than 40000 product reviews from 25 different domains [26]. This database was also used by[27-29].

The trained Naïve Bayesian classifier chosen [30] provides probability of the input text being either positive or negative. However, the sentiment can have three classes: positive; negative; and neutral. Hence, an additional post-processing step is needed to firmly decide which class the input text belongs to. Beside investigating the use of a threshold approach similar to [25], to determine the class, this paper proposes, investigates and compares the use of two clustering methods to cluster the probabilities to three clusters. Since the number of clusters is known, the two clustering techniques investigated here are the k-means and k-medoid algorithms [31, 32].

IV. Experiments and Discussions

This section describes a set of experiments conducted to analyze properties of the video product reviews and examine the performance of the proposed system with different suggested post-processing algorithms.

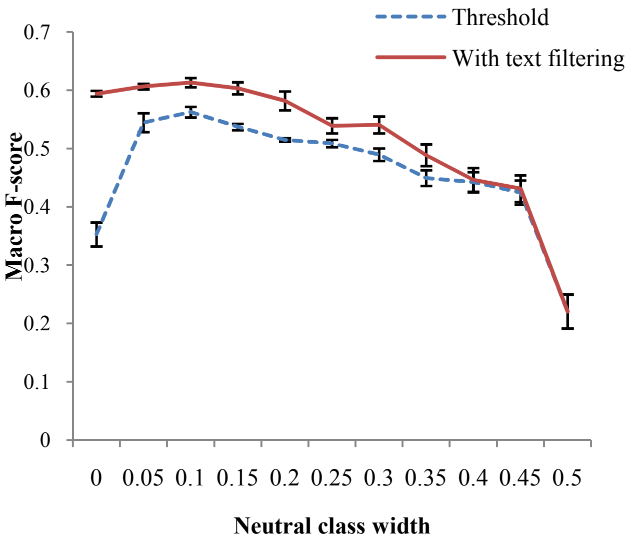

Fig. 2. The macro F-score of the sentiment classification system for different range of the neutral class with 0.05 confidence error bars.

A. *Test Data*

A set of 50 English review video clips about 5 different products (10 videos per product) from 5 different domains was downloaded from YouTube in November 2012. The test set was prepared manually since to the authors extent of knowledge there is a lack of relevant test corpuses for user video blogs [3, 10]. The collected videos were published by reviewers from different genders, countries and ethnicities. The total duration of the videos was 61.2 minutes with an average of 1.224 minutes per video. The five products are: "Ipad mini" (computer); "Norton antivirus" (software); "Galaxy S3" (phone); Coriolanus (movie); and Nivea (beauty cream). Also included were the recent user comments associated with each video. The video text transcription used for the sentiment classification system of Section III was generated using the YouTube video transcriber.

The video database contains 7765 words constituting 522 statements (sentences). The word error rate of the YouTube transcriber was manually calculated to be 25.2%, which is considered high compared with the latest NIST evaluation [24]. Also, the accuracy of the subtitle displaying time was 100% which means that any correctly detected sentiment will always be associated to the correct time in the video. The sentiment of each sentence was manually labeled by an expert human listener as ground truth (as in [1, 2, 6, 25]) on the basis that statements containing positive or negative opinion words or phrases about the target product are labeled as positive or negative, respectively. Statements of facts about the product or containing opinions about another product (in the case of product comparisons) were labeled as neutral. This produced 163 positive, 101 negative and 258 neutral statements, respectively which is the real size of the test data since the target is classifying the statements not the videos. Since no previous system is known to solve the same problem the system will be compared with the result of the sentiment classifier when the input is manually entered and error free as ground truth.

B. *Propeirties of User Generated Video Reviews*

After analyzing the downloaded videos it was found that they have the same properties found in [10], providing broad context for deeper understanding of the product. It was found that in 96% of the videos, a detailed evaluative description with visual demonstration of the product was provided. This visual demonstration makes the camera most of the time is focusing on the product not in the reviewers face. This has the effect of distracting the sentiment detection by emotion like the one used in [11] because of the lack of facial expressions. Also, in 94% of the videos another product from the same domain was mentioned or compared with the target product, which makes them more attractive and understandable. From 1291 text comments, only 243 comments (18.82%) were related to the product and the remaining were unrelated, being about the presentation quality or the author of the video. In addition, 8% of these videos contained no comments. Because of the previous problems with comments and despite that they are written manually and do not suffer from recognition error, it has been decided to analyze the video transcription produced by an ASR. As mentioned earlier the ASR used is the YouTube transcriber which did recognize 5805 word correctly from the 7765 of the test database which is 74.76% accuracy.

C. *Sentiment Classification Performance*

To evaluate the proposed system, the F-score of classifying the 522 sentences to the three classes was calculated since it is the common sentiment evaluation measure [16, 33-35]. Specifically, the macro-averaged F-score was chosen to give equal weight to the three classes despite population differences [33]. In this section the effect of applying a fixed threshold on the Naïve classifier output probabilities and the alternative of using clustering will be presented and compared. In addition the effect of using text filtering will also be investigated.

1) *Fixed Threshold*

The dashed curve in Figure 2 shows the macro F-score of the system for different values of the classification threshold (the solid curve will be explained later). The classification threshold here is half of the values range considered, as Neutral starting from the center (a probability of 0.5). This means that at point range (neutral class width) of 0.05, for example, only statements that receive positive and negative scores between 0.45 and 0.55 will be classified as neutral, with statements having positive scores above 0.55 considered positive and remaining probabilities indicating negative statements. It can be seen that when the neutral class width is zero (only sentences with probability of 0.5 considered neutral) the F-score is 0.35. The F-score rises with increasing neutral class width until it reaches a maximum of 0.56 at a width of 0.1 (neutral range from 0.4 to 0.6), since more correct neutrals were in this range. The F-score then reduces gradually because more positive and negative statements begin to enter the neutral class range.

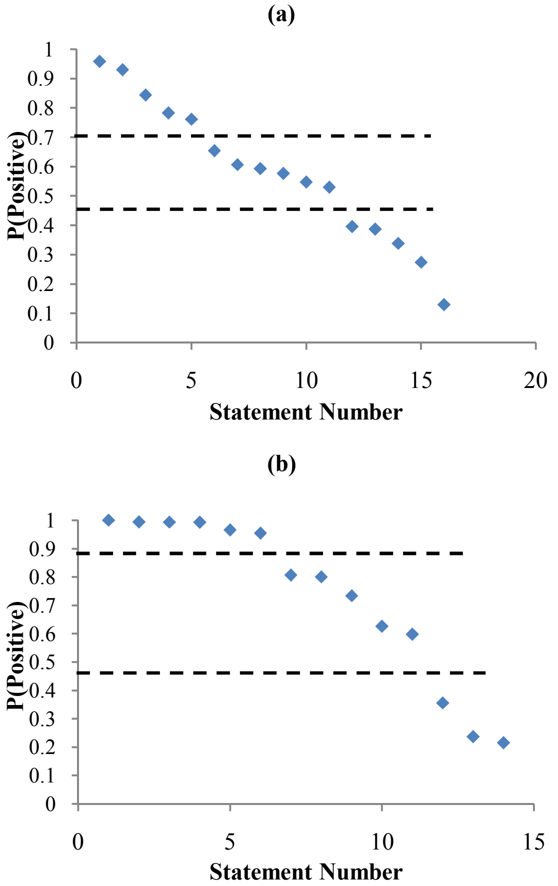

Fig. 3. Scatters of statements sentiment probabilities extracted from two different videos by two different bloggers

2) Probabilities Clustering

A static threshold for all videos was found to be unsuitable because different speakers and videos tend to use different vocabulary and opinion words that affect the Naïve classifier probabilities (some speakers use very firm emotive words (e.g. swearing) and others do not). Also speakers have different accents, which result in different word error rates for the speech recognition process. Figure 3 shows a scatter of the sentiment probabilities for each statement for two different videos for two different speakers. It can be seen from the dashed lines that the class borders (thresholds) should be different for the two videos to obtain higher accuracy. In video (b) the speaker was more firm, resulting in positive sentiment probabilities above 0.9 and the WER for this video was 29% while speaker (a) was not as firm and the speech WER was 37.1%.

To overcome these problems, this paper proposes clustering of the output probabilities from the Naïve classifier to adaptively choose the most appropriate cluster borders for each input video. This paper investigated the k-mean and k-medoid clustering algorithms [31, 32]. When k-means was applied instead of the fixed threshold the macro F-score was 0.6 while the k-medoid gave a macro F-score of 0.61. The k-medoid uses an actual data point as a cluster center rather than an average as in k-means. Both clustering methods produce similar F-scores that are significantly better than using the fixed threshold.

TABLE I. MACRO F-SCORES WHEN APPLYING THE ADJUSTED THRESHOLD, K-MEAN AND K-MEDOID CLUSTERING WITH AND WITHOUT TEXT FILTERING

Method	Threshold	k-means	k-medoid
NO Text Filtering	0.56	0.6	0.61
Text Filtering	0.61	0.65	0.66

3) Text Filtering

As mentioned in earlier, it is very common in video reviews to find opinionated sentences about another product compared with the target one. The opinionated words in such sentences causes their Naïve Bayesian output probabilities to be biased to positive or negative while they should be neutral since they are about another product. So, it is proposed in this paper to filter the video text transcription first and introduce the sentences that contain the target product name or a pronoun ("it", "its", "it's" and "this") that is assumed to be referring to the target product to the Naïve classifier. The remaining sentences are then added to the neutral class. Results show an average improvement in F-score of the system with the suggested text filtering compared to no text filtering as can be seen in the solid curve of Figure 2.

Table I summarizes the results by listing the system's macro F-score using the adjusted threshold, k-means and k-medoid clustering with and without applying the text filtering. It can be seen that text filtering improves the results for all three methods and that while the k-medoid is slightly better than the k-means, both clustering methods performed better than the adjusted threshold. The proposed system can achieve a macro F-score of 0.66. This is very promising knowing that the used Naïve classifier gives 0.78 macro F-score when correct manually written text (that has no recognition error) is used [30].

V. CONCLUSIONS AND FUTURE WORK

This paper investigated temporal sentiment detection for user generated video product reviews. The paper proposed a system that combines ASR and a Naïve Bayesian classifier and then processes the output probabilities to produce a final decision. It was found that clustering the Naïve classifier output probabilities results in F-scores significantly higher than using a fixed threshold. It was also found that assigning sentences that do not contain the target product name or a relevant pronoun to the neutral class significantly improves the results. The system achieves an F-score of 0.66, which is very promising considering that the classifier gives an F-score of 0.78 when the input text from the ASR system is error free. From the presented results it is believed that analyzing automatically generated video text transcription for temporal sentiment detection is feasible and deserves future research. In the future, extracting emotion from the speech tone and its combination with the text sentiment classifier will be investigated.

ACKNOWLEDGMENT

This work is supported by the Smart Services CRC, Sydney, Australia.

References

[1] Z. Kunpeng, C. Yu, X. Yusheng, D. Honbo, A. Agrawal, D. Palsetia, K. Lee, L. Wei-keng, and A. Choudhary, "SES: Sentiment Elicitation System for Social Media Data," in Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, 2011, pp. 129-136.

[2] S. Choudhury and J. G. Breslin, "User sentiment detection: a YouTube use case," presented at the The 21st National Conference on Artificial Intelligence and Cognitive Science, 2010.

[3] M. M. S. Missen, M. Boughanem, and G. Cabanac, "Opinion Detection in Blogs: What Is Still Missing?," presented at the Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining, 2010.

[4] H. J. Cheong and M. A. Morrison, "Consumers' reliance on product information and recommendations found in UGC," Journal of Interactive Advertising, vol. 8, pp. 38-49, 2008.

[5] X. Dou, J. A. Walden, S. Lee, and J. Y. Lee, "Does source matter? Examining source effects in online product reviews," Computers in Human Behavior, Elsevier, pp. 1555-1563, 2012.

[6] D. A. Ostrowski, "Sentiment Mining within Social Media for Topic Identification," in IEEE Fourth International Conference on Semantic Computing (ICSC), 2010, pp. 394-401.

[7] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in Mining Text Data, ed: Springer, 2012, pp. 415-463.

[8] P. Melville, W. Gryc, and R. D. Lawrence, "Sentiment analysis of blogs by combining lexical knowledge with text classification," presented at the Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris, France, 2009.

[9] Z. Xiaoyu, X. Changsheng, C. Jian, L. Hanqing, and M. Songde, "Effective Annotation and Search for Video Blogs with Integration of Context and Content Analysis," IEEE Transactions on Multimedia Tools and Applications, vol. 11, pp. 272-285, 2009.

[10] K. Wilson, "Crowd-Sourced Evaluation: A Qualitative Study of User-Generated Product Review Videos on ExpoTV. com," Journal of MultiDisciplinary Evaluation, vol. 8, 2012.

[11] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in Proceedings of the 13th international conference on multimodal interfaces, 2011, pp. 169-176.

[12] M. S. Barakat, C. H. Ritz, and D. A. Stirling, "An Improved Template-Based Approach to Keyword Spotting Applied to the Spoken Content of User Generated Video Blogs," in IEEE International Conference on Multimedia and Expo (ICME), 2012, pp. 723-728.

[13] 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, 2004, pp. 168-177.

[14] S. M. Kim and E. Hovy, "Determining the sentiment of opinions," in Proceedings of the 20th international conference on Computational Linguistics, 2004, p. 1367.

[15] K. Durant and M. Smith, "Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection," Advances in Web Mining and Web Usage Analysis, pp. 187-206, 2007.

[16] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," in Proceedings of LREC, 2010, pp. 1320-1326.

[17] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2(1-2), 2008.

[18] K. Harrenstien. (2009). The Official Google Blog. Available: http://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html

[19] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices," in Proc. ICASSP, 2006, pp. 185-188.

[20] Pocketsphinx software package. (2011). avilable at: http://sourceforge.net/projects/cmusphinx/. Available: http://sourceforge.net/projects/cmusphinx/

[21] M.S.Barakat, C.H.Ritz, and D.A.Stirling, "Detecting Offensive User Video Blogs: An Adaptive Keyword Spotting Approach " in International Conference of Audio, Language and Image Processing (ICALIP), 2012, pp. 419-425.

[22] D. Brezeale and D. J. Cook, "Automatic Video Classification: A Survey of the Literature," IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 38, pp. 416-430, 2008.

[23] M. Nakamura, K. Iwano, and S. Furui, "Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance," Computer Speech and Language, vol. 22, pp. 171-184, 2008.

[24] NIST. (2010). National Institute of Standards and Technology, ASR History. Available: http://www.itl.nist.gov/iad/mig/publications/ASRhistory/index.html

[25] J. Kågström, "Improving naive bayesian spam filtering," Master thesis, Mid Sweden University, 2005.

[26] M. Dredze. (2009). Multi-Domain Sentiment Dataset (version 2.0). Available: http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

[27] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman, "Learning bounds for domain adaptation," Advances in neural information processing systems, vol. 20, pp. 129-136, 2007.

[28] M. Dredze, K. Crammer, and F. Pereira, "Confidence-weighted linear classification," in Proceedings of the 25th international conference on Machine learning, 2008, pp. 264-271.

[29] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation with multiple sources," Advances in neural information processing systems, vol. 21, pp. 1041-1048, 2009.

[30] uClassify. Sentiment. Available: http://www.uclassify.com/browse/uclassify/Sentiment

[31] G. Peters, M. Lampart, and R. Weber, "Evolutionary Rough k-Medoid Clustering Transactions on Rough Sets VIII." vol. 5084, J. Peters and A. Skowron, Eds., ed: Springer Berlin / Heidelberg, 2008, pp. 289-306.

[32] A. Reynolds, G. Richards, and V. Rayward-Smith, " Intelligent Data Engineering and Automated Learning – IDEAL 2004." vol. 3177, Z. Yang, et al., Eds., ed: Springer Berlin / Heidelberg, 2004, pp. 173-178.

[33] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach," Journal of Informetrics, vol. 3, pp. 143-157, 2009.

[34] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis," Computational linguistics, vol. 35, pp. 399-433, 2009.

[35] H. Kanayama and T. Nasukawa, "Fully automatic lexicon expansion for domain-oriented sentiment analysis," in Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 2006, pp. 355-363.