A Novel Approach for Emotion Classification based on Fusion of Text and Speech

Ali Houjeij, Layla Hamieh, Nader Mehdi, Hazem Hajj
Department of Electrical and Computer Engineering
American University of Beirut
Beirut, Lebanon
{akh11, lsh04, nam22, hh63}@aub.edu.lb

Abstract— In this paper we design a system that adopts a novel approach for emotional classification from human dialogue based on text and speech context. Our main objective is to boost the accuracy of speech emotional classification by accounting for the features extracted from the spoken text. The proposed system concatenates text and speech features and feeds them as one input to the classifier. The work builds on past research on music mood classification based on the combination of lyrics and audio features. The innovation in our approach is in the specific application of text and speech fusion for emotion classification and in the choice of features, Furthermore, in the absence of benchmark data, a dataset of movie quotes was developed for testing of emotional classification and future benchmarking. The comparison of the results obtained in each case shows that the hybrid text-speech approach achieves better accuracy than speech or text mining alone.

Keyword: Algorithms, emotional classification, speech mining, text mining, fusion.

I. INTRODUCTION

The tremendous growth and availability of data on the internet has motivated researchers to extract useful information from this data. In fact, researchers have been particularly interested in employing emotional classification with text and speech mining in an attempt to render human-computer interaction more natural where speech may replace traditional input devices. Computers can then react to affect cues in verbal content [1][2], and not just understand it. This has been of great interest for the human–machine interaction research community. It is argued that one of the most important human skills a computer should acquire to understand the human is the affective aspect of communication [2]. Many interesting applications exist in this domain such as spoken dialogue systems in call-center applications and automated telephone systems to detect customer dissatisfaction [1][3]. Authors who have tackled the particular issue of music mood classification based on both audio and lyrics have also pointed out its importance in organizing and accessing music digital libraries [4][5]. This improvement in content-based information retrieval is not only limited to audio databases but extends to search engines and text retrieval technologies. In addition to boosting machine intelligence, emotional classification and in particular that associated with text mining has been found beneficial in analyzing computer-mediated communication (e.g. chat servers, discussion forums, blogs…). The authors in [6] highlight the benefits of extracting information from chat messages in certain areas like financial forensics and threat analysis. Analogous to emotional classification, opinion mining is equally beneficial in decision support in various areas such as shopping, entertainment, government, research and development, marketing and education [7].

Taking into consideration the increasing interest of data mining research in exploiting machine learning techniques for emotional classification, we present in our work an improvement in the accuracy of emotional classification by combining the features of audio and text. The work builds on past research [4,5,14,15] on music mood classification based on the combination of lyrics and audio features. We aim at analyzing a speaker's speech and the corresponding text content in order to identify the speaker's emotional state. Accordingly, each user will be assigned one if the following states: angry, afraid, happy, sad or neutral. Each of the text and speech will be processed separately and classified before undergoing a fusion technique where the features of both are treated as one set used for classification. The uniqueness of the system is reflected in using the hybrid approach in emotion classification for human speech. This allows the comparison of both approaches in emotional classification, text and speech mining, with the hybrid approach especially in terms of accuracy.

The rest of the paper is organized as follows. In section II, we present three streams of related work: emotional classification in text, emotional classification in speech and fusion approaches in emotional classification using both text and speech mining. Section III describes the general architecture of our emotional classification system. We describe the experiments performed and the results obtained along with an analysis of these results in section IV. Finally, we conclude our work in section V highlighting directions for future work.

II. RELATED WORK

In this section, we present a review of three different areas of related work: emotional classification in text, emotional classification in speech and fusion approaches in emotional classification using both text and speech mining.

978-1-4673-0747-5/12/$31.00 ©2012 IEEE

A. *Emotional classification from text*

Researchers have been increasingly showing interest in employing emotional classification with text mining and have addressed various approaches in different contexts. In [8], Li et al. introduced an approach for text sentiment analysis in the context of detecting online hotspot forums. Their work, which targeted 31 different online sports forums, was divided into two main stages: emotional polarity computation and sentiment analysis. Their analysis was based on support vector machine (SVM) and k-means clustering. In the first stage the authors calculated a value for each post in the forum that measures the polarity and intensity of the emotions in this post. In the second stage these values were used along with other attributes related to each forum (total number of posts, number of responses to each post…) to cluster the available forums. The centers of the obtained clusters were then identified as the hotspot forums. The results obtained by k-means clustering were used as training data for the SVM classifier which was then applied for new forums for hotspot identification. Based on accuracy, sensitivity and prediction values, the authors compared the results of both techniques and concluded that they produce consistent results. Another example of mood classification in online texts is addressed in the work of Mishne [9] which explores blog posts. Mishne used a large variety of text features: frequency counts of words, length of posts, emotional polarity (negative or positive) of posts, Pointwise Mutual Information (PMI) which gives numerical weights to keywords based on how related each word is to a certain mood, emphasized words and special symbols (emoticons and punctuation marks). SVM was used for classification but the success rates achieved in his experiments were relatively low mainly because the annotation of posts used in the training sets was determined by the authors themselves and was thus subject to their own definition of emotions. Also, the styles of the authors were diverse and inconsistent and the classification was affected by the short length of blog entries. Kucukyilmaz et al. [6] explored text mining in subjective computer-mediated communication to predict attributes related to users (such as gender and age) and messages (such as receiver and time). In their work, they used two competing approaches based on the type of features: style-based approach which exploits stylistic features such as word lengths and punctuation usage and term-based approach which exploits the vocabulary used in the messages. Del-Hoyo et al. [2], however, concentrated less on the context in which to apply emotional classification and more on the approach followed in this classification. They explored three different approaches: the statistical approach, the semantic approach and a hybrid statistical-semantic system. In the statistical approach, the features used were the Term-Frequency-Inverse Document Frequency (TF-IDF) values of the terms in the text. The TF-IDF is a measure of the frequency of a term in a certain document as well as the fraction of documents in which it occurs. Combining the statistical and semantic features for classification showed a 13% increase in classification accuracy over the semantic approach and a 3% increase in accuracy over the statistical technique.

Based on what has preceded, SVM was the most popular classifier used in text mining and thus we will be using it in our work.

B. *Emotional classification from speech*

Emotional classification from speech is still a hot research area and many novel techniques are introduced and published every year. Ververidis and Kotropoulos [10] provided a detailed literature survey of all the resources, features and methods used in emotional speech recognition. In their work, they discussed that speech data used in the developed systems is either natural (spontaneous), simulated (acted professionally) or elicited (neither natural nor simulated). They presented a summary of the various features and classification techniques used by researchers in this domain. In [3], Vayranen et al. described a new technique for emotional classification from speech in spoken Finnish using vowel-length segments. They use simulated emotional speech examples to train and test the proposed scheme. For each audio sample, they collected vocal and local prosodic features of the audio to obtain a large feature set. The vocal features capture the intensity of the speech while the prosodic are those related to the syllable length and pitch of speech sounds. For each subset (local and vocal), KNN (k-nearest neighbors) classifier was used to classify the data and predict the emotions. Then, a decision level fusion technique that uses a weighed sum of classifier scores (probability estimates of the kNN classifiers) was applied to combine both results. The results showed that the average accuracy of the fusion technique was better than each classifier alone and better than the human guessing. In their work, Grimm et al. [11] introduced an evaluation of emotions in speech. They presented a three dimensional emotional space composed of emotion primitives (Dominance, Valence and Activation). All other emotions were considered linear combinations of these primitives. In their proposed system, the audio segments were preprocessed and then pitch, speaking rate, intensity and spectral related features were extracted. A fuzzy logic estimator was then used on the above features to estimate the emotion primitives. Finally, they used KNN classifier to map emotional primitives to emotional categories. The simulation results showed that the above system accomplished an overall recognition rate of 85 %. Iliev et al. developed a new method for speech emotional recognition based on glottal airflow signals (pressure on the mouth) [12]. The authors argued that the glottal flow is highly affected by the emotional state since it represents the degree of tension on the vocal folds. They derived a mathematical model for the speech production in order to extract the glottal features. Several classification algorithms were tested, and the experimental results verified that Optimum-path forest (OPF) [13] performed the best. This scheme is hard to apply in real time classification since it needs complex instruments (accurate speakers attached to the neck) to capture the glottal air flow signal.

C. Hybrid approach

Combining text and speech mining for emotional classification was mostly explored in the field of music genre classification where lyric and audio features were combined to improve the classifier's performance. Researchers addressed two issues: (1) the effect of hybrid systems on the performance of the classifier and (2) the best fusion technique to use in combining audio and lyric.

In [4], [5], [14] and [15] experiments showed that hybrid systems using early and late fusion techniques had an absolute advantage over individual algorithms where they produced higher classification accuracy for all mood categories. The Work in this domain addresses different areas (music mood, music genre in digital libraries…). Most of it compared two fusion techniques; the first one used separate predictions for audio and lyrics and combined them through voting, while the second exploited interdependencies between aspects from both modalities combined the audio and lyric features into one vector for classification. In [4] and [16] the first fusion technique produced better results while in [5], [14] and [15] the second one performed better. Some other factors affect the extent to which algorithms based on combined features improve the classifier's performance. For example, in [4] and [16] the choice of the features to combine was based on their effect in the individual algorithm (audio or text based) and in [5] the number of dimensions taken in the lyric or audio space affected the performance. The authors in [4] shed the light on an additional advantage of hybrid systems which is the ability of these systems to achieve similar or higher accuracies than audio or text based algorithms using smaller number of training sets.

All the work discussed above used a hybrid approach based on lyrics and audio features for music mood classification. We will use a similar approach to combine the audio and textual features of a human speech in order to enhance the accuracy of available emotion classification systems that only consider speech or text feature.

III. PROBLEM DESCRIPTION

Machines are still unable to understand human's expressions and emotions. One way to feed this information to computers and extract it from humans is through analyzing the human's voice and the spoken text. In the following section, we present a fusion (text-speech) method to solve the problem of human emotional classification. Our proposed method introduces the hybrid text-speech approach for emotional classification in the context of human speech in contrary to the past approaches which concentrated on music mood classification using audio and lyrics features. The aim is to boost the performance of emotional classification of human speech by accounting for the features of the spoken text. For this purpose, additional audio features were introduced and new text features were derived. Furthermore, in the absence of benchmark data, a dataset of movie quotes was developed for testing of emotional classification and future benchmarking.

IV. METHODOLOGY

Our approach aims at using audio and textual features in order to identify the emotions embedded in a certain speech which belongs to one of the following emotional classes: angry, afraid, happy, sad and neutral. These emotional classes are the most basic and common classes of human emotions. The main steps of the approach are data preprocessing, feature extraction and finally classification. These steps are discussed in further detail in the sections below.

A. Data set collection

Our collected data set contains 350 different movie quotes along with their corresponding texts extracted online [17]. The movies from which the quotes were extracted were chosen on the basis of making our dataset diverse enough to have the five different emotional tags equally expressed. The movies were carefully selected in a way to represent all emotions without any bias. A movie was a good candidate if the subject was comedy, love, misery, horror or a war movie. Comedy and love movies contain happy emotions, misery movies contain sad emotions, horror movies contain fear and war movies contain angry emotions. Neutral emotions were easy to find as they appeared in many normal narrations or conversations. The chosen quotes were usually small in size, ranging between one to two sentences. Having short segments helped in assigning the emotions since the actor does not change the mood in the same segment. The speech and text quotes were collected and stored as audio wave and text files respectively. Four different people (2 Male and 2 Female) listened to these quotes and assigned an emotional tag for each of them. We assumed that four listeners were enough to make sure of the emotional tag of a certain segment. If all listeners agreed on the same emotional tag, then this meant that the audio segment has a clear emotional tag and hence it was used. Else, the segment was marked as vague in order not to use it in training the classifier. This process was important to ensure that the data was tagged correctly and to remove all instances of inaccurate data. The training database was then purified where all quotes with vague emotions were removed from the database. This reduced the size of the database to 262 quotes. The annotated data was used for developing the classification models and testing their accuracies.

B. Data preprocessing and feature extraction

In the second stage, the generated data set was preprocessed to extract features from both text and speech to be input to the SVM classifier.

1) Text preprocessing and feature extraction

In the text mining part of our system, we used the semantic approach focusing on the emotional information of each word in the sentence using part of speech tagging and standard lexicons. The details are described below.

a) Data Preprocessing

To obtain a set of representative terms for each text, we started by eliminating all non-alphanumeric characters that were not numbers or letters. Then we performed stopwords filtering

since these terms appeared frequently in all texts and were thus non-discriminating features. Since there is no standard stopwords list for English language, we used one of the lists available online at webconfs.com [18] which consists of the most common stopwords ignored by search engines. The following step was part of speech tagging where only verbs, nouns, adjectives and adverbs were preserved. For this purpose, we used the Stanford part of speech tagger [19].

b) Feature Extraction

To account for the significance of emotional information in keywords [2], we used the English lexical database WordNet Affect [19] to obtain emotional tags. For each text, a score Xi for each of the emotional states i, where i ranged from 1 to 5 accounting for all five emotional states considered in our system. The vector X containing the Xi elements represented the feature vector.

Computing the score Xi was conducted as follows: after obtaining the part of speech, WordNet [20] was used to determine the set of synonyms for each word (synset). Since WordNet Affect assigns an affective label to each word, it was used to assign an emotional tag for each synonym in the synset. The most common emotion (sad for example) in the synset was the most probable emotion for this word and therefore the entry X_j in the X feature vector corresponding to that most probable emotion (sad) was incremented. This process was repeated for all the words in a certain text, resulting in the final feature vector X for the text. Figure 1 illustrates how the vector X is computed for a certain text A.

Algorithm: *Computing the feature vector*

Input
- A: Text to be mined

Output
- X: Feature vector

```
(1)  X[1..5]=0
(2)  KeyWords=filter(A)
(3)  for word w in KeyWords
(4)       syn=synset(w)
(5)       temp[1..5]=0
(6)       for synonym s in syn
(7)            temp[emotion(s)]++
(8)       end for
(9)       X[maxIndex(temp)]++
(10) end for
```

Figure 1. Algorithm for text mining

2) Speech preprocessing and feature extraction

In order to extract the emotional state from speech, we first preprocessed and cleaned the collected audio segments and then extracted the relevant features from them.

a) Data Preprocessing

The frequency of human voice ranges between 300 and 3400 Hz [20].Consequently, to only keep data corresponding to the human voice and remove all irrelevant noise data, a bandpass filter with cutoff frequencies [300 Hz, 3400 Hz] was applied to each audio segment.

b) Feature Extraction

The emotional state of the speaker directly influences many diverse features of the speech. For example, the volume of the sound of an angry speaker is different from that of a calm speaker. Similarly, pitch and speaking rate change radically with emotional state. In our system, we used 5 different types of speech related features in order to diversify the features:

Spectral:

- **Audio Spectrum rolloff** is the frequency below which 85% of the spectral energy lies.
- **Audio Spectrum centroid** is the center of mass of the spectrum. It is the mean of the frequencies of the signal weighted by their magnitudes, determined using a Fourier transform.
- **Mel-Frequency Cepstral Coefficients** (MFCCs) which express the audio signal on a mel frequency Scale (linear below 1000Hz and logarithmic above 1000Hz). These coefficients help in identifying phonetic characteristics of speech.

Temporal:

- **Zero Crossing** In a given period, the number of times the time domain signal crosses zero is the zero crossing.
- **Log-attack time** is the time taken by a signal to reach its maximum amplitude from a minimum threshold time.

Various: **Audio Spectral Flatness features** measure the amount of peaks in the power spectrum of an audio signal to identify noise-like sounds since white noise is characterized by high spectral flatness.

C. Classification

1) Text and Speech Classification

The data vectors collected from speech and text were separately input to the SVM classifier. 10-fold cross validation was used to test the accuracy of each alone. These tests were used as base to compare with the proposed fusion method.

2) Fusion Technique

In our method, audio and textual features were consolidated into a single feature vector before being input to the classifier (SVM). Consequently, multiple feature sets were integrated in

order to generate a single classifier. This method, as expected [5], exploits the interdependencies between the different features being combined by feeding the joint data to the same classifier. This is central in our case because our data set consists of written and spoken versions of the same text, thus the features extracted from each are interdependent. Hence, this technique leverages the complementary information extracted from audio and text.

V. Experiments and Results

In order to test the effectiveness of the algorithm, three experiments were conducted for detecting which emotional classification algorithm provided the highest accuracy. In the first experiment we tested for speech classification. First, we converted all media files to wav (Waveform Audio File Format) since it has a better quality and it is easier to interact with MATLAB. The SVM classifier was trained with the collected 262 vectors each with 183 attributes. The 183 attributes were the speech extracted features described in the section IV. In the second experiment we calculated the accuracy of a text emotional classification algorithm. Again, the SVM classifier was trained with 262 vectors each with 5 attributes representing the score of each emotion. In the third experiment we calculated the accuracy of the fusion algorithm. This time, the SVM classifier was trained with 262 vectors each with 188 attributes. Both speech and text extracted features were concatenated and used as an input to the classifier. In the three experiments the classification was done with 10-fold cross validation. The results of the experiments are shown in Table 1.

Table1 shows that all three algorithms provided a better estimate than a random guess (20%). The text mining didn't do well since the emotional tags of the texts were speech and context dependent, i.e. the training data was labeled after hearing the speech and reading the text. This affects the accuracy of text mining alone since it can't capture the voice while the labels do. As an example, a movie quote from GodFather "They killed my son" should be classified as a sad statement. But since the existing text mining approach does not detect a word with sad sentiment, it does not classify the text as sad. It is worth noting that this presents an opportunity for future research on improved semantic extraction that can infer context and emotions even in the absence of affective words. The hybrid approach was able to achieve the highest accuracy measure among others. It provides an excellent improvement over text classification and good improvement over speech, though not as significant. This is the case since the hybrid approach exploits the relationships between the speech and the corresponding text thus achieving a better clustering for the data set and consequently a better accuracy in the whole system. This proves that combining text and audio features can enhance the accuracy of emotional classification of human speech. At first, the fusion approach wasn't so effective since the number of added features (5) was small compared to the original (180). So, to investigate this more, we carefully chose 50 of the 180 features and did another analysis for the speech classification and the fusion technique. Note that the results were more impressive for the fusion technique as it was able to provide a good improvement. Finally, the accuracies obtained are comparable to those obtained in the literature. For example, [3] achieved lower accuracy, [8,9] achieved similar results while [2,12] achieved higher accuracies.

Algorithm	Accuracy
Speech emotional classification (180 features)	45.42%
Text emotional classification (5 features)	26.36%
Fusion approach (180+5 features)	46.57 %
Speech emotional classification (50 features)	35.88%
Fusion approach (50+5 features)	31.30%

Table 1 Simulation results

VI. Conclusion and Future Work

In this paper we introduced a new emotional classification method for human speech. This method is based on exploiting both the audio and the corresponding textual features. The method introduces new speech features and new text features into account. The speech features are a large combination of various speech characteristics. The text features depend on the most probable emotion in the word's synset. The algorithm produced higher accuracies than existing systems that exploit only a single type of feature. These findings can help improve effectiveness of emotion classification of human speech and thus make human machine interaction a more realistic subject. Furthermore, a data set was generated for use in benchmarking emotional classification from audio and text.

As a direction of future work, we are investigating the effect of choosing the appropriate number of features used in speech classification in order to enhance the overall performance of the system. One way is to try different combinations of speech features and compare the obtained accuracies. Improved text mining methods will be considered including semantic and sentiment inferences. Also, the above algorithm can be used along with speech to text software to boost the performance of speech mining.

ACKNOWLEDGMENT

This work was partially funded by the KASCT-Intel's Middle East Energy Efficiency Research (MER) and the American University of Beirut (AUB). Special acknowledgment is due to Ms. Lama Nachman for her valuable insights.

REFERENCES

[1] T. Vogt, E. Andre, and J. Wagner, "Automatic Recognition of Emotions from Speech: a Review of the Literature and Recommendations for Practical Realisation," Affect and Emotion in Human-Computer Interaction, pp.75–91,2008.

[2] R. del Hoyo, I. Hupont, F. Lacueva, and D. Abad´ıa, "Hybrid Text Affect Sensing System for Emotional Language Analysis," in Proceedings of the ACM International Workshop on Affective-Aware Virtual Agents and Social Robots2009, pp. 1–4.

[3] E. Vayrynen, J. Toivanen, and T. Seppanen, "Classification of Emotion in Spoken Finnish Using Vowel Length Segments: Increasing Reliability with a Fusion Technique," Speech Communication, 2010 vol. 53, no.3, pp. 269-282.

[4] X. Hu and J. Downie, "Improving Mood Classification in Music Digital Libraries by Combining Lyrics and Audio," in Proceedings of the ACM 10th annual joint conference on Digital libraries. 2010, pp. 159–168.

[5] C. Laurier, J. Grivolla, and P. Herrera, "Multimodal Music Mood Classification Using Audio and Lyrics," Seventh International Conference on Machine Learning and Applications (ICMLA'08) 2008, pp. 688–693.

[6] T. Kucukyilmaz, B. Cambazoglu, C. Aykanat, and F. Can, "Chat Mining: Predicting User and Message Attributes in Computer-Mediated Communication," Information Processing & Management, vol. 44, no. 4, pp. 1448–1466, 2008.

[7] H. Binali, V. Potdar, and C. Wu, "A State of the Art Opinion Mining and its Application Domains," in Industrial Technology, 2009. ICIT 2009. IEEE International Conference. 2009, pp. 1–6.

[8] N. Li and D. Wu, "Using Text Mining and Sentiment Analysis for Online Forums Hotspot Detection and Forecast," Decision Support Systems, vol. 48, no. 2, pp. 354–368, 2010.

[9] G. Mishne, "Experiments with Mood Classification in Blog Posts," in Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access. 2005.

[10] D. Ververidis and C. Kotropoulos, "Emotional Speech Recognition: Resources, Features, and Methods," Speech Communication, vol. 48, no. 9, pp. 1162–1181, 2006.

[11] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," Speech Communication, vol. 49, no. 10-11, pp. 787–800, 2007.

[12] A. Iliev, M. Scordilis, J. Papa, and A. Falc˜ao, "Spoken Emotion Recognition Through Optimum-Path Forest Classification Using Glottal Features," Computer Speech & Language, vol. 24, no. 3, pp. 445–460, 2010.

[13] J.P. Papa, A.X. Falcao, and C.T.N. Suzuki. "Supervised Pattern Classification Based on Optimum-Path Forest". International Journal of Imaging Systems and Technology, vol.19 no.2pp.120-131, 2009.

[14] R. Neumayer and A. Rauber, "Integration of Text and Audio Features for Genre Classification in Music Information Retrieval," Advances in Information Retrieval, pp. 724–727, 2007.

[15] Y. Yang, Y. Lin, H. Cheng, I. Liao, Y. Ho, and H. Chen, "Toward Multi-Modal Music Emotion Classification," Advances in Multimedia Information Processing-PCM2008,pp.70–79,2008.

[16] X. Hu, J. Downie, and A. Ehmann, "Lyric Text Mining in Music Mood Classification," American music, vol. 183, no. 5,049, pp. 2–209, 2009.

[17] Wave Planet, Online Audio Database http://www.wavplanet.com Retrieved on Feb, 2011.

[18] WebConfs. Stop words. http://www.webconfs.com/stop-words.php,2006.

[19] Kristina Toutanova and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.

[20] C. Strapparava and A. Valitutti, "WordNet-Affect: an Affective Extension of WordNet," in Proceedings of LREC, vol. 4. Citeseer, 2004, pp. 1083–1086.

[21] A. Norrell and D. Walsh, "Method and Apparatus for Adaptively Equalizing a Signal Received from a Remote Transmitter," Mar. 7 2000, US Patent 6,034,993.

[22] H. Gunes and M. Piccardi, "Affect Recognition from Face and Body: Early Fusion vs. Late Fusion," 2005 IEEE International Conference in Systems, Man and Cybernetics, vol. 4. 2006, pp. 3437–3443.