2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science (IRI) | 978-1-6654-6603-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/IRI54793.2022.00052

Audio Sentiment Analysis using Spectrogram and Bag-of-Visual-Words

Sophina Luitel
Department of Computer Science
North Carolina Agricultural and Technical State University
Greensboro, USA
sluitel@aggies.ncat.edu

Mohd Anwar
Department of Computer Science
North Carolina Agricultural and Technical State University
Greensboro, USA
manwar@ncat.edu

***Abstract*—Audio sentiment analysis has many applications in a present-day context, such as call center environments, conversational agents, and human-robot interactions. However, analyzing sentiment using audio signals is a significant challenge due to the difficulty of accurately determining the robust feature set needed to detect sentiments expressed within the audio signal. It is novel to use spectrogram and bag-of-visual-words for representing robust audio features for sentiment analysis. Therefore, we propose using keypoints of the spectrogram (a 2D image representation of frequencies in a signal) to classify sentiment. We converted audio signals to spectrograms using Short Time Fourier Transform (STFT). The Oriented FAST and Rotated BRIEF (ORB) algorithm extracted the salient keypoints scattered on a regular grid over the spectrogram. Then, using the Bag-of-Visual-Words (BoVW) technique, the descriptors from each column of the spectrogram of the audio signal are converted to histograms. We have trained and tested models using a multilingual dataset. We have applied hyperparameter tuning while generating the histograms to increase the accuracy of the model. These histograms are then passed to classifiers, and the sentiment of each audio signal is classified. Our experimental result shows that Random Forest was the best classifier with an accuracy of 76% and an F1 score of 78% and demonstrates the prospect of our approach in language-agnostic audio sentiment analysis.**

***Index Terms*—Sentiment Analysis, Spectrograms, Bag-of-Visual-Words, Multi-lingual Sentiment Analysis, Audio Pattern Recognition, Machine Learning.**

I. Introduction

Research in the sentiment analysis area is expanding due to its broad application. We can see its application in business, education, politics, and finance [1]; essentially everywhere that human interaction is involved. This explains the extensive attention this research field is getting. Furthermore, since humans display their emotion through spoken voice, facial expression, and written text, sentiment analysis is being done utilizing all these display streams together or individually. There is a plethora of research on text-based sentiment analysis [2] and recently, there has been an increasing trend in using multimodal-based analysis [3]. Previously, audio sentiment analysis used automatic speech recognition (ASR) tools to transcribe audio recordings and then applied text-based analysis [4] but recently, much research has used the raw audio and extracted the acoustic features [5]. The use of spectrograms can also be found in the audio-based sentiment analysis field due to the advancements of computer vision technology [6]. Moreover, the visual-based sentiment analysis is also rising due to the advancement in computer vision technology and proliferation of visual online content. Deep learning has made facial expression recognition the most trending research field in computer vision [7]. In this paper, we propose a method inspired by the research done by authors in [8]. We have utilized spectrogram, the visual representation of the audio, and used open-source feature extractor ORB in our proposed method instead of a patented extractor Speeded Up Robust Features (SURF) that previous researchers were using. We wanted to determine whether the open-source tools can give comparable results to a patented extractor. Additionally, we have applied hyperparameter optimization, where we tried different sample rates of creating spectrogram and a different number of clusters (k in k-means clustering algorithm) to show how we could increase the accuracy of the model. We built 10 different models to get the best combination of hyperparameters. We used a small dataset to propose this method and experimented with hyperparameter tuning to present the accuracy increment of the model. The spectrogram representation of audio allows sentiment analysis to be language independent. We have used audio recordings in three languages: English, Italian, and Spanish to prove that our model is language independent, adding to the research area of language-independent sentiment analysis. To the best of our knowledge, this is the first time that this combination of techniques (spectrogram and bag-of-visual words) is used in this multilingual dataset to build a language-agnostic sentiment analysis model. We recognize that the language-independent model could have many benefits like various linguistic resources are not available for many languages [9]. This kind of language-agnostic model could benefit in such cases where there are no training datasets for a particular language. Additionally, it would reduce the computational cost because a single model would suffice for sentiment analysis for multiple languages.

Our paper adds to the work of a language-independent audio sentiment analysis approach. We presented a simple yet effective method of using spectrograms as an intermediate data representation of audio for sentiment analysis. We used and

978-1-6654-6603-5/22/$31.00 ©2022 IEEE
DOI 10.1109/IRI54793.2022.00052

validated the use of open-source keypoints extractor algorithm ORB (Oriented Fast and Rotated Brief) over patented extractor tools. The model was built using simple histograms from spectrograms to classify the sentiment of the audio files using SVM and Random Forest algorithm-based classifiers, where hyperparameter optimization improved the model's accuracy from 66% to 76% i.e., by 15%.

Our main contributions are the following.

1) We compare the current state-of-the-art audio sentiment analysis techniques and determine their strengths and limitations.
2) We build language-agnostic models for sentiment analysis using three different languages: English, Italian and Spanish.
3) We ascertain the efficacy of open-source keypoint extractor ORB over patented extractors like Scale-Invariant Feature Transform (SIFT) and Speeded Up Robust Features (SURF), getting reasonable results.
4) We develop a pipeline for language-agnostic audio sentiment analysis with a composition of spectrogram, ORB and bag-of-visual words.
5) We test the impact of hyperparameter optimization in improving models' accuracy. We applied combinations of different sample rates and different numbers of clusters in k-means clustering algorithms to improve the models' accuracy.

The rest of the paper is organized as follows. Section II discusses the literature review, highlighting the past research by dividing it into text-based and audio feature-based sentiment analysis sections. In section III, we describe our methodology and the components used. We provide a description of the dataset in section IV. Section V discusses the environment and parameters used in the experiment, and in section VI, we present our results. Finally, we conclude and present our future work in section VII.

II. Related Works

There are many approaches that researchers have explored for sentiment analysis from human speech. For example, linguistic and non-linguistic approaches to audio signals have been used for sentiment analysis. The widely used methods include transcribing the text, hand-crafted feature extraction from audio signals, multimodal analysis, and using deep neural networks for automatic feature selection. Based on literature review, audio sentiment analysis methods can be divided into two types: text-based and audio feature-based. Audio feature-based sentiment analysis focuses on the features extracted from audio such as intensity, loudness, Mel-frequency cepstral coefficients, and pitch followed by the analysis on the extracted features. Besides text-based and audio feature-based sentiment analysis, we propose using the raw audio waveforms as the input for generating a 2D image (spectrogram) to be used for sentiment analysis.

We propose to use spectrograms as an intermediate representation for the audio sentiment analysis because it is language independent. Due to the advancement of image processing fields, transfer learning from the image domain to audio could be beneficial. Below we have discussed current state-of-the-art techniques that have been used so far for audio sentiment analysis and in Table I we have listed their strengths and limitations.

A. Text-based approach

A sentiment lexicon is a list of lexical features (e.g., words) which are generally labeled corresponding to their semantic orientation as either positive or negative [10]. Commonly used lexicons include binary polarity-based lexicons such as Harvard General Inquirer [11] and Linguistic Inquiry and Word Count (LIWC) [12]. Liu compared the vector space language model such as Word2Vec, Global Vectors for Word Representation (GloVe) for sentiment analysis using Reddit data [13]. Ezzat et al. used text as an intermediate representation of speech by converting speech to text for sentiment analysis using automatic speech recognition (ASR) tools to transcribe audio to text and using different text classification and clustering techniques for the end goal of sentiment analysis [4].

- Limitations of text-based sentiment analysis:
    1) The approach causes word-sense disambiguation (i.e., which sense of a word is used) due to multiple meanings associated with them.
    2) The approach is language dependent.
    3) An extra step is required to convert the speech to text, and only then the analysis can be done.

B. Audio feature-based approach

In some approaches, audio recordings were converted to text using automated speech recognition (ASR) tools; however, much research has adopted a model that focuses not on what was said but on how it was said – the acoustic aspect of speech [14]. Furthermore, some approaches extracted the signal features and used multimodal features as well. We have categorized the literature on audio feature-based sentiment analysis into the following sections.

1) Sentiment Analysis using hand-crafted features: Hand-crafted features have been used in many other domains along with sentiment analysis. One of the research papers highlighted the strength of this approach. Authors showed that using hand-crafted features outperformed the CNN across smaller sample sizes and with increased interpretability for classifying liver Magnetic Resonance Imaging MRI adequacy [15]. However, there are some limitations we could find in this approach. Kishore et al. extracted Mel-frequency cepstral coefficients (MFCCs) and wavelet features and used Gaussian Mixture Models. They indicated models with wavelet features performed better compared to MFCC features however, none produced satisfactory results. They concluded that the performance could be improved by fusion of features extracted from multimodal data such as speech data or face data [16]. Tickle et al. studied features extraction using the openSMILE tool from the speech wave and reduced the count of features to more manageable features and used a supervised neural network for

sentiment analysis. Among the 988 acoustic features extracted using the openSMILE tools only 71 were used. They used WEKA machine learning tool where "info-gain attribute" and "classify Subset Eval" algorithms were used to rank the features in the order of importance [17]. Likewise, selective acoustic features such as MFCCs or other low-level descriptors (pitch, energy, filterbank features) were extracted from a raw speech for emotion recognition in Augilar et al. [18]. It was a challenging task because selecting a suitable feature set is not easy, whereas the approach we propose does not need a manual selection of features; instead, distinct keypoints from spectrograms are selected using the ORB algorithm.

- Limitations of sentiment analysis using hand-crafted features:
  1) It is a challenge to select the suitable feature set.
  2) The approach requires a lot of effort, i.e., domain knowledge, labor and time, to design hand-crafted features

2) Sentiment analysis using multimodal features: There is an increase in research concerning multimodal sentiment analysis rather than individual modalities such as text, audio, or visual [19] [20] [21]. Authors in [22] presented a method for multimodal sentiment classification based on the utterance-level visual data stream. They concluded that using multiple modalities could lead to an error rate reduction of up to 10.5% compared to one modality at a time. Multimodal-based sentiment analysis is getting attention because of the plethora of multimodality sources of information online, such as YouTube (videos and audio) [23]. Li et al. combined the lexical and acoustic model with the late fusion approach [24]. OpenSMILE was used to extract acoustic features and lexical aspects captured by backoff n-gram language models. These models were fused for sentiment analysis. A bimodal (linguistic and acoustic) approach was proposed for sentiment analysis by [25], and their results showed that relying on the joint use of both modalities gave better performance for sentiment analysis than using only one modality at a time.

- Limitation of sentiment analysis using multimodal features:
  1) Complex approach that requires proper fusion techniques.

C. Our Approach

We propose using spectrograms for audio sentiment analysis. Spectrograms have been used for other audio analysis tasks such as content classification [26] and stress recognition [27]. Negi et al. plotted a spectrogram for each audio file with the feature extracted (MFCC) using Python Library (Librosa), which were then fed to a convolutional neural network (CNN) for depression detection [28]. Spyrou et al. used a bag-of-visual-words methods adopted from the computer vision [8]. They used the Speeded-Up Robust Features (SURF) method for visual feature extraction from critical points of spectrograms, created clusters (BoVW), and then built a histogram which was passed to the Support Vector Machine (SVM) classifier. We have used the similar idea for *language-agnostic audio sentiment analysis* by using Bag-of-Visual-Words together with open-source keypoint extractor algorithm and performed the hyperparameter optimization to increase the model's accuracy.

- Limitation of sentiment analysis using spectrogram:
  1) The approach is still at developmental stage for sentiment analysis.
  2) The approach requires a large set of datasets for training deep neural networks.

III. Dataset

We have used the EmoFilm dataset [29], a multilingual corpus of emotional speech, consisting of 1115 English, Spanish, and Italian emotional utterance extracted from 43 films and 207 speakers (both male and female are included). This dataset with the combination of spectrograms and bag-of-visual-words technique has never been used before for sentiment analysis. Table II shows all five emotions that were incorporated with the file count of each emotion. The label of each audio file was in the filename, which made it very easy to get the label of emotion depicted in the audio signal (i.e., Fear, Disgust, Happiness, Anger, Sadness). We have chosen Happiness and Sadness for our binary sentiment classification purpose. We chose only female audio files for happiness and sadness to minimize the variance. All three languages were included in this experiment to build language-independent models.

IV. Methodology

We develop a pipeline for spectrogram-based audio sentiment analysis displayed in Fig 1. We have discussed the steps of the pipeline in the following subsections.

A. Spectrogram Creation

We used Short-Time Fourier Transform to obtain a two-dimensional spectrogram using the Python package (librosa). The y-axis was converted to a log scale, and the color dimension was converted to decibels. Fig 2 shows spectrograms of happy sentiment on the left and sad on the right side, created at 22,050 Hz sample rate. The spectrograms were resized to 240X240 pixels. We created spectrograms of three different variations of sample rates: 22,050 Hz, 44,100 Hz, and 66,150 Hz.

B. Key-points Extraction

We extracted keypoints from the spectrogram with 32-bit descriptors and built a visual dictionary using the ORB algorithm. Keypoints are the stand out points in an image; so whether the image rotates, shrinks, or expands, its keypoints will always be the same. Moreover, descriptors are the description of the keypoints. ORB was developed at OpenCV labs [30] in 2011 as an efficient and viable alternative to SIFT and SURF. ORB became popular mainly because SIFT and SURF are patented algorithms; however, ORB is free to use. ORB builds on the well-known features from the Accelerated and Segments Test (FAST) keypoint detector and the Binary Robust Independent Elementary Feature (BRIEF) descriptor.

TABLE I
STRENGTHS AND LIMITATIONS OF CURRENT STATE-OF-THE-ART TECHNIQUES FOR AUDIO SENTIMENT ANALYSIS

Method	Strengths	Limitations
Text-based sentiment analysis	• Plethora of past research, tools, and techniques	• Causes word-sense disambiguation (i.e., which sense of a word is used) due to multiple meanings associated with them • Language-dependent approach • Requires an extra step to convert the speech to text
Sentiment analysis using hand-crafted features	• Provides interpretable results • Works better with a small set of data	• A challenge to select the suitable feature set • Requires a lot of effort, i.e., domain knowledge, labor and time to design hand-crafted features
Sentiment analysis using multimodal features	• Using multiple modalities could lead to an error rate reduction of up to 10.5% compared to one modality at a time. • Large volume of multimodality sources of information available online such as, YouTube	• Complex approach: must build model for each modality • Requires proper fusion techniques that may not always be achieved
Sentiment analysis using spectrogram	• Language-independent approach • Transfer learning is applicable from image processing domain	• Still at developmental stage for sentiment analysis • Requires a large set of datasets for training deep neural networks

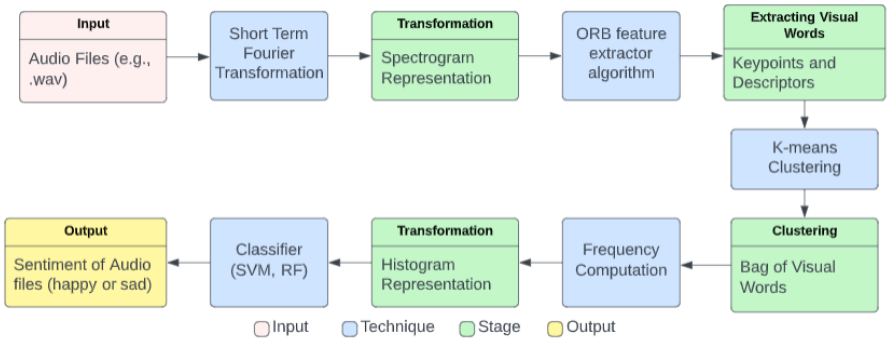

Fig. 1. Pipeline for the proposed spectrogram-based audio sentiment analysis

TABLE II
EMOFILMS DATASET

Emotion	Emotion Label	Count of Audio Files
Fear	ans	221
Disgust	dis	168
Happiness	**gio**	**240**
Anger	rab	232
Sadness	**tri**	**254**

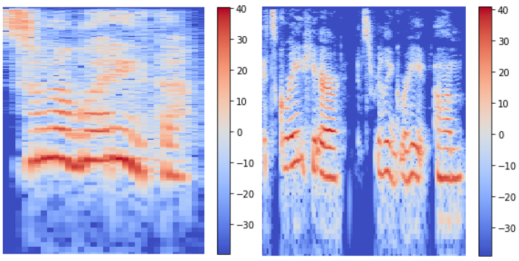

Fig. 2. Spectrograms of happy and sad sentiment

C. Bag-of-Visual-Words (BOVW) Creation

BOVW is commonly used in image classification [31], and the concept is adapted from information retrieval and NLP's bag of words (BOW). BOVW is weakly supervised model, built on the idea of visual vocabularies [8]. Instead of counting words that appear in a document, in BOVW, we use image features (key points and descriptors) as words and make a frequency histogram.

D. Clustering

After extracting the descriptors from each image, we made clusters from the descriptors using K-means clustering. We created multiple clusters (5, 10, 20) for each image to compare and get the best performance. It is a part of hyperparameter optimization. The center of each cluster is used as the visual dictionary's vocabularies.

E. Histogram

We made a frequency histogram from vocabularies for each image. Those histograms are our Bag-of-Visual-Words, saved as a .csv file. We generated 5 different .csv files of the histogram data; three files with cluster size of 5 (k=5) and three different sample rates (22,050 Hz, 44,100 Hz, and 66,150 Hz), and two other files with a sample rate of 22,050 Hz and cluster size of 10 and 20. Finally, we passed the histogram files to classifiers (Random Forest and Support Vector Machines). We have labels of sentiment from the filename itself which made it easier to extract the label for the audio file. We experimented to see if the frequency of clusters appearing in the image could help classify the sentiment(happiness vs sadness).

TABLE III
NUMBER OF MODELS USED IN THE EXPERIMENT

Model	Sample Rate (Hz)	Number of Cluster	Classifier
1	22,050	5	RF
2	44,100	5	RF
3	66,150	5	RF
4	22,050	5	SVM
5	44,100	5	SVM
6	66,150	5	SVM
7	22,050	10	RF
8	22,050	20	RF
9	22,050	10	SVM
10	22,050	20	SVM

V. EXPERIMENT

We used Python in Google Colab Pro to run our experiments. We ran our experiment in two parts. First, we generated the histogram using 200 audio files (100 happy and 100 sad) and saved it as a .csv file. Then, we used the .csv file as the input for the classifiers. In the first part, we used the Librosa library to generate spectrograms (200X200-pixel size) from audio files. We generated spectrograms with different sample rates (22,050 Hz, 44,100 Hz, and 66,150 Hz) for comparison. Then, we used ORB algorithms to extract keypoints with their descriptors. We converted the RGB image to grayscale before extracting the keypoints. K-means clustering algorithm was then used to get cluster sizes of 5, 10, and 20 and generated histograms to get the Bag-Of-Visual Dictionary. We created 5 different files and saved it as .csv for the next part of the experiment. We built 10 different models using two classifiers (SVM-based and RF-based), three sample rates (22,050 Hz, 44,100 Hz, 66,150 Hz), and three variations of clusters(5, 10, 20) to get the best model. Table III shows our 10 different models. After transforming the raw audio dataset into the histogram .csv files, we split the dataset into the ratio of 75:25 for training and testing. We used the training data to train our classifier models. We used Random Forest and SVM classifier. When we created histograms from 5 clusters and 22,050 Hz sample rates, we got 66% accuracy from the Random Forest model. We applied RandomizedSearchCV with 3 folds of cross-validation for our hyperparameter tuning in Random Forest Classifier, and we were able to optimize our model to get 74% accuracy. Our best estimators were 'bootstrap': False, 'criterion': 'entropy', 'max_depth': 7, 'max_features': 'log2', 'min_samples_leaf': 44, 'min_samples_split': 44, 'n_estimators': 1133.

For SVM, we used GridSearchCV for hyperparameter tuning. Before hyperparameter tuning, the accuracy was 67%, and after the tuning, we were able to get 73% accuracy. The best estimators were C=1, gamma=0.001, kernel='poly'. We got the best accuracy of 76% for the Random Forest and 73% for the SVM classifier. We compared the result of different combinations of clusters of descriptors and sample rates and found that for our dataset, the best combination was 10 cluster size and 22,050 Hz sample rate.

TABLE IV
MODEL ACCURACY WITH 5 CLUSTERS AND DIFFERENT SAMPLE RATES

Sample Rate(Hz)	RF	SVM
22,050	74%	73%
44,100	70%	64%
66,150	54%	56%

TABLE V
MODEL ACCURACY WITH DIFFERENT CLUSTERS AND 22,050 HZ SAMPLE RATES

Clusters	RF	SVM
5	74%	73%
10	76%	72%
20	74%	72%

VI. RESULTS

The classifiers were trained using keypoints from spectrograms with three different sample rates and histograms of cluster size of 5. The accuracies of each classifier for detecting happiness vs. sadness are shown in Table 4. The RF classifier has achieved accuracies of 74%, 70%, and 54% with spectrograms' sample rates of 22,050 Hz, 44,100 Hz and, 66,150 Hz respectively. The SVM-based classifier produced 73%, 64% and 56% accuracy with sample rates 22,050 Hz, 44,100 Hz and, 66,150 Hz respectively. Once we got the best sample rates for both classifiers, we fixed that sample rate of 22,050 Hz and tried using a different number of clusters (k). For hyperparameter optimization, we have compared 3 different cluster sizes of descriptors (5, 10, 20) and 3 different sample rates for spectrograms (22,050 Hz, 44100 Hz, and 66,150 Hz). Table 5 shows the accuracy of the models with three different cluster sizes. We found that with 10 clusters at a sample rate of 22,050 Hz, Random Forest classifier performs best at 76% accuracy without any hyperparameter tuning, and for SVM, it performs best at 73% accuracy with 5 clusters at a sample rate of 22,050 Hz. Table 6 displays the F1 score, recall, and precision percentage for 2 different classifiers with the combination that performed the best. The F1 score, recall, and precision of Random Forest classifier with 10 clusters and 22,050 Hz sample rate were 78%, 75%, and 81% respectively. In the SVM model with 5 clusters and a 22,050 Hz sample rate, the F1 score was 68%, a recall was 56%, and precision was 87%.

TABLE VI
F1 SCORE, RECALL AND PRECISION FOR BEST OF THE RF AND SVM MODELS WITH 22,050 HZ SAMPLE RATE

	RF (10 clusters)	SVM (5 Clusters)
F1 Score	78%	68%
Recall	75%	56%
Precision	81%	87%

VII. CONCLUSION AND FUTURE WORK

There are different approaches to audio sentiment analysis, and we compared them for strength and limitations. We

proposed a novel simple approach for audio sentiment analysis using spectrogram of audio/speech and bag-of-visual-words methods. We built language agnostic models using a dataset of three different languages: English, Italian, and Spanish. We believe this could be of an important addition to language-independent audio sentiment analysis research. We validated the use of open-source keypoint feature extraction tool - ORB, in lieu of patented tools such as SIFT/SURF, which has been used in past research. We generated spectrogram from the audio recordings, used ORB and k-means clustering algorithms to generate histograms and finally used SVM and Random Forest classifiers for sentiment classification. This combination of techniques has never been used before. We optimize hyper-parameters through generating spectrogram (different sample rates) and histogram (different number of clusters) to increase the models' accuracy. The Random Forest classifier achieved the best accuracy of 76% which outperforms state-of-the-art spectrogram sentiment analysis [6].

We performed binary classification (happy or sad) for this experiment; however, to fully explore the potential of this method, we plan to apply this approach for multiclass classification in the future. Additionally, we used a small dataset and two well-known machine learning algorithms to demonstrate the prospect of this approach. In the future, we plan to use deep learning algorithms with larger datasets to improve the accuracy of the proposed method. Moreover, we plan to test the languages that were not in the training set to determine the versatility of the approach.

REFERENCES

[1] D. Alessia, F. Ferri, P. Grifoni, and T. Guzzo, "Approaches, tools and applications for sentiment analysis implementation," *International Journal of Computer Applications*, vol. 125, no. 3, 2015.

[2] H. P. Patil and M. Atique, "Sentiment analysis for social media: a survey," in *2015 2nd International Conference on Information Science and Security (ICISS)*. IEEE, 2015, pp. 1–4.

[3] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.

[4] S. Ezzat, N. El Gayar, and M. M. Ghanem, "Sentiment analysis of call centre audio conversations using text classification," *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 4, no. 1, pp. 619–627, 2012.

[5] F. Mairesse, J. Polifroni, and G. Di Fabbrizio, "Can prosody inform sentiment analysis? experiments on short spoken reviews," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 5093–5096.

[6] G. Pikramenos, G. Smyrnis, I. Vernikos, T. Konidaris, E. Spyrou, and S. J. Perantonis, "Sentiment analysis from sound spectrograms via soft bovw and temporal structure modelling." in *ICPRAM*, 2020, pp. 361–369.

[7] K. Patel, D. Mehta, C. Mistry, R. Gupta, S. Tanwar, N. Kumar, and M. Alazab, "Facial sentiment analysis using ai techniques: state-of-the-art, taxonomies, and challenges," *IEEE Access*, vol. 8, pp. 90 495–90 519, 2020.

[8] E. Spyrou, R. Nikopoulou, I. Vernikos, and P. Mylonas, "Emotion recognition from speech using the bag-of-visual words on audio segment spectrograms," *Technologies*, vol. 7, no. 1, p. 20, 2019.

[9] M. Shams, N. Khoshavi, and A. Baraani-Dastjerdi, "Lisa: language-independent method for aspect-based sentiment analysis," *IEEE Access*, vol. 8, pp. 31 034–31 044, 2020.

[10] B. Liu *et al.*, "Sentiment analysis and subjectivity." *Handbook of natural language processing*, vol. 2, no. 2010, pp. 627–666, 2010.

[11] P. J. Stone, D. C. Dunphy, and M. S. Smith, "The general inquirer: A computer approach to content analysis." 1966.

[12] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," Tech. Rep., 2015.

[13] Y. Liu, "A comparative study of vector space language models for sentiment analysis using reddit data," Ph.D. dissertation, North Carolina Agricultural and Technical State University, 2020.

[14] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, 2015.

[15] W. Lin, K. Hasenstab, G. Moura Cunha, and A. Schwartzman, "Comparison of handcrafted features and convolutional neural networks for liver mr image adequacy assessment," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.

[16] K. K. Kishore and P. K. Satish, "Emotion recognition in speech using mfcc and wavelet features," in *2013 3rd IEEE International Advance Computing Conference (IACC)*. IEEE, 2013, pp. 842–847.

[17] A. Tickle, S. Raghu, and M. Elshaw, "Emotional recognition from the speech signal for a virtual education agent," in *Journal of Physics: Conference Series*, vol. 450, no. 1. IOP Publishing, 2013, p. 012053.

[18] G. Aguilar, V. Rozgić, W. Wang, and C. Wang, "Multimodal and multi-view models for emotion recognition," *arXiv preprint arXiv:1906.10198*, 2019.

[19] L. C. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in *Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat.*, vol. 1. IEEE, 1997, pp. 397–401.

[20] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2008.

[21] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling," in *Proc. INTERSPEECH 2010, Makuhari, Japan*, 2010, pp. 2362–2365.

[22] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance-level multimodal sentiment analysis," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 973–982.

[23] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, 2016.

[24] B. Li, D. Dimitriadis, and A. Stolcke, "Acoustic and lexical sentiment analysis for customer service calls," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5876–5880.

[25] Y. Jia and S. SungChu, "A deep learning system for sentiment analysis of service calls," *arXiv preprint arXiv:2004.10320*, 2020.

[26] H. Lee, P. Pham, Y. Largman, and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," *Advances in neural information processing systems*, vol. 22, 2009.

[27] L. He, M. Lech, N. Maddage, and N. Allen, "Stress and emotion recognition using log-gabor filter analysis of speech spectrograms," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–6.

[28] H. Negi, T. Bhola, M. S Pillai, and D. Kumar, "A novel approach for depression detection using audio sentiment analysis," *International Journal of Information Systems & Management Science*, vol. 1, no. 1, 2018.

[29] E. Parada-Cabaleiro, G. Costantini, A. Batliner, A. Baird, and B. Schuller, "Categorical vs dimensional perception of italian emotional speech," 2018.

[30] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[31] J. Sivic and A. Zisserman, "Efficient visual search of videos cast as text retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 4, pp. 591–606, 2008.