A Deep Learning System for Sentiment Analysis of Service Calls

Yanan Jia
Businessolver / Bellevue, WA
yjia@businessolver.com

Abstract

Sentiment analysis is crucial for the advancement of artificial intelligence (AI). Sentiment understanding can help AI to replicate human language and discourse. Studying the formation and response of sentiment state from well-trained Customer Service Representatives (CSRs) can help make the interaction between humans and AI more intelligent.

In this paper, a sentiment analysis pipeline is first carried out with respect to real-world multi-party conversations - that is, service calls. Based on the acoustic and linguistic features extracted from the source information, a novel aggregated method for voice sentiment recognition framework is built. Each party's sentiment pattern during the communication is investigated along with the interaction sentiment pattern between all parties.

1 Introduction

The natural reference for AI systems is human behavior. In human social life, emotional intelligence is important for successful and effective communication. Humans have the natural ability to comprehend and react to the emotions of their communication partners through vocal and facial expressions (Kotti and Paternò, 2012; Poria et al., 2014a). A long-standing goal of AI has been to create affective agents that can recognize, interpret and express emotions.

Early-stage research in affective computing and sentiment analysis has mainly focused on understanding affect towards entities such as movie, product, service, candidacy, organization, action and so on in monologues, which involves only one person's opinion. However, with the advent of Human-Robot Interaction (HRI) such as voice assistants and customer service chatbots, researchers have started to build empathetic dialogue systems to improve the overall HRI experience by adapting to customers' sentiment.

Sentiment study of Human-Human Interactions (HHI) can help machines identify and react to human non-verbal communication which makes the HRI experience more natural. The call center is a rich resource of communication data. A large number of calls are recorded daily in order to assess the quality of interactions between CSRs and customers. Learning the sentiment expressions from well-trained CSRs during communication can help AI understand not only what the user says, but also how he/she says it so that the interaction feels more human.

In this paper, we target and use real-world data - service calls, which poses additional challenges with respect to the artificial datasets that have been typically used in the past in multimodal sentiment researches (Cambria et al., 2017), such as variability and noises. The basic 'sentiment' can be described on a scale of approval or disapproval, good or bad, positive or negative, and termed polarity (Poria et al., 2014b).

In the service industry, the key task is to enhance the quality of services by identifying issues that may be caused by systems of rules, or service qualities. These issues are usually expressed by a caller's anger or disappointment on a call. In addition, service chatbots are widely used to answer customer calls. If customers get angry during HRI, the system should be able to transfer the customers to a live agent. In this study, we mainly focuses on identifying 'negative' sentiment, especially 'angry' customers. Given the non-homogeneous nature of full call recordings, which typically include a mixture of negative, and nonnegative statements, sentiment analysis is addressed at the sentence level. Call segments are explored in both acoustic and linguistic modalities. The temporal sentiment patterns between customers and CSRs appearing in calls are described.

The paper is organized as follows: Section 2 covers a brief literature review on sentiment recognition

Proceedings of the 3rd Workshop on e-Commerce and NLP (ECNLP 3), pages 24–34
Online, July 10, 2020. ©2020 Association for Computational Linguistics

from different modalities; Section 3 proposes a pipeline which features our novelties in training data creation using real-world multi-party conversations, including a description of the data acquisition, speaker diarization, transcription, and semi-supervised learning annotation; the methodologies for acoustic and linguistic sentiment analysis are presented in Section 4; Section 5 illustrates the methodologies adopted for fusing different modalities; Section 6 presents experimental results including the evaluation measures and temporal sentiment patterns; finally, Section 7 concludes the paper and outlines future work.

2 Related Work

In this section, we provide a brief overview of related work about text-based and audio-based sentiment analysis.

2.1 Text-based Sentiment Analysis

Sentiment analysis has focused primarily on the processing of text and mainly consists of either rule-based classifiers that make use of large sentiment lexicons, or data-driven methods that assume the availability of a large annotated corpora.

Sentiment lexicon is a list of lexical features (e.g. words) which are generally labeled according to their semantic orientation as either positive or negative (Liu, 2010). Widely used lexicons include binary polarity-based lexicons, such as Harvard General Inquirer (Stone et al., 1966), Linguistic Inquiry and Word Count (LIWC, pronounced 'Luke') (Pennebaker et al., 2007, 2001), Bing (Liu, 2012), and valence-based lexicons, such as AFINN (Nielsen, 2011), SentiWordNet (Alhazmi et al., 2013), and SnticNet (Cambria et al., 2010). Employing these lexical, researchers can apply their own rules or use existing rule-based modeling, such as VADER (Hutto and Gilbert, 2015), to do sentiment analysis. One big advantage for the rule-based models is that these approaches require no training data and generalize to multiple domains. However, since words are annotated based on their context-free semantic orientation, word-sense disambiguation (Hutto and Gilbert, 2015) may occur when the word has multiple meanings. For example, words like 'defeated', 'envious', and 'stunned' are classified as 'positive' in Bing, but '-2' (negative) in AFINN. Although the rule-based algorithm is known to be noisy and limited, a sentiment lexicon is a useful component for any sophisticated sentiment detection algorithm and is one of the main resources to start from (Poria et al., 2014b).

Another major line of work in sentiment analysis consists of data-driven methods based on a large dataset annotated for polarity. The most widely used datasets include the MPQA corpus which is a collection of manually annotated news articles (Wiebe et al., 2005; Wilson et al., 2005), movie reviews with two polarity (Pang and Lee, 2004a), a collection of newspaper headlines annotated for polarity (Strapparava and Mihalcea, 2007). With a large annotated datasets, supervised classifiers have been applied (Go et al., 2009; Pang and Lee, 2004b; dos Santos and Gatti, 2014; Socher et al., 2013; Wang et al., 2016). Such approaches step away from blind use of keywords and word co-occurrence count, but rather rely on the implicit features associated with large semantic knowledge bases (Cambria et al., 2015).

2.2 Audio-based Sentiment Analysis

Vocal expression is a primary carrier of affective signals in human communication. Speech as signals contains several features that can extract linguistic, speaker-specific information, and emotional. Related work about audio-based sentiment analysis along with multimodal fusion is reviewed in this section.

Studies on speech-based sentiment analysis have focused on identifying relevant acoustic features. Use open source software such as OpenEAR (Eyben et al., 2009), openSMILE (Eyben et al., 2010), JAudio toolkit (McEnnis et al., 2005) or library packages (McFee et al., 2015; Sueur et al., 2008) to extract features. These features along with some of their statistical derivates are closely related to the vocal prosodic characteristics, such as a tone, a volume, a pitch, an intonation, an inflection, a duration, etc.

Supervised or unsupervised classifiers can be fitted based on the statistical derivates of these features (Jain et al., 2018; Pan et al., 2012). Sequence models can be fitted based on filter banks, Mel-frequency cepstral coefficients (MFCCs), or other low-level descriptors extracted from raw speech without feature engineering (Aguilar et al., 2019). However, this approach usually requires highly efficient computation and large annotated audio files.

Multimodal sentiment analysis has started to draw attention recently because of the unlimited multimodality source of information online, such as videos and audios (Cambria et al., 2017; Poria, 2016; Poria et al., 2015). Most of the multimodal sentiment analysis is focused on monologue videos. In the last few years, sentiment recognition in conversations has started to gain research interest, since reproducing human interaction requires a deep understanding of conversations, and sentiment plays a pivotal role in conversations. The existing conversation datasets are usually recorded in a controlled environment, such as a lab, and segmented into utterances, transcribe to text and annotated with emotion or sentiment labels manually. Widely used datasets include AMI Meeting Corpus (Carletta et al., 2006), IEMOCAP (Busso et al., 2008), SEMAINE (Mckeown et al., 2013) and AVEC (Schuller et al., 2012).

Recently, a few recurrent neural network (RNN) models are developed for emotion detection in conversations, e.g. DialogueRNN (Majumder et al., 2019) or ICON(Hazarika et al., 2018). However they are less accurate in emotion detection for the utterances with emotional shift (Poria et al., 2019) and the training data requires the speaker information. The conversation models are not employed in our polarity sentiment analysis because of the quality of the data and the approach used to gain the training data. More detailed explanations can be found in Section 3.4.

At the heart of any multimodal sentiment analysis engine is the multimodal fusion (Shan et al., 2007; Zeng et al., 2007). The multimodal fusion integrates all single modalities into a combined single representation. Features are extracted from each modality of the data independently. Decision-level fusion feeds the features of each modality into separate classifiers and then combines their decisions. Feature-level fusion concatenates the feature vectors obtained from all modalities and feeds the resulting long vector into a supervised classifier. Recent research on multimodal fusion for sentiment recognition has been conducted at either the feature level or decision level (Poria, 2016; Poria et al., 2015).

3 Dataset and Pipeline

The data resources used for our experiments are described in Section 3.1. Data preparation including speech transcription and speaker diarization is discussed in Section 3.2. The sentiment annotation guideline is introduced in Section 3.3. Section 3.4 presents a semi-supervised learning annotation pipeline that chains data preparation, model training, model deploying and data monitor.

3.1 BSCD: Benefits Service Call Dataset

The main dataset we created in this paper consists of service calls collected from a health care benefits Call Center (named BSCD). Calls are focused on customers looking for help or support with company provided benefits such as health insurance.

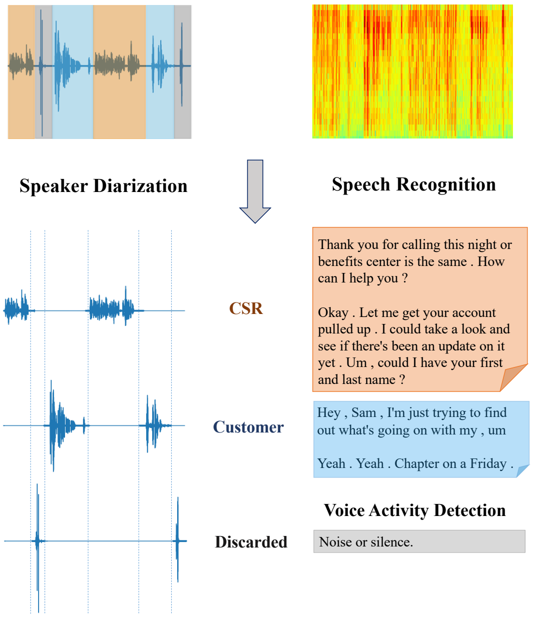
500 calls are collected from the call center database covering diverse topics, such as insurance plan information, insurance id card, dependent coverage, etc. The call dataset has female and male speakers randomly selected with their age ranging approximately from 16-80. Calls involving translators are eliminated to keep only speakers expressing themselves in English. All the calls are presented in Wave format with a sample rate of 8000 Hertz and duration varying from 4 minutes to 26 minutes. All calls are pre-processed to eliminate repetitive introductions. The beginning of each call contains an introduction of the users' company name by a robot. To address this issue, the segment before the first pause (silence duration $>$ 1 second) is removed from each call.

A robust computational model of sentiment analysis needs to be able to handle real-world variability and noises. While the previous researches on multimodal sentiment or emotion analysis use audio and visual recorded in laboratory settings (Busso et al., 2008; Mckeown et al., 2010, 2013); the BSCD gathers real-world calls which contain ambient noise present in most audio recordings, as well as diversity in person-to-person communication patterns. Both of these conditions result in difficulties that need to be addressed in order to effectively extract useful data from these sources.

3.2 Data Preparation

To discard noise and long pauses (silence duration $>$ 5 seconds) in calls, Voice Activity Detection (VAD) is applied, followed by the application of Automatic Speech Recognition (ASR) and Automatic Speaker Diarization (ASD) to transcribe the verbal statements, extract the start and end time of each utterance, and identify the speaker of each utterance. Each call is segmented into an average of 69 utterances. The duration of the utterances is right-skewed with a median of 2.9 seconds; first and third quantiles 1.6 and 5.1 seconds.

By searching keywords such as 'How can I help' in the content of each utterance, speakers are labeled



Thank you for calling this night or benefits center is the same . How can I help you ?

Okay . Let me get your account pulled up . I could take a look and see if there's been an update on it yet . Um , could I have your first and last name ?

Hey , Sam , I'm just trying to find out what's going on with my , um

Yeah . Yeah . Chapter on a Friday .

Noise or silence.

Figure 1: Data preparation workflow

as CSR or customer. Each utterance is linked to the corresponding audio stream, auto transcription, as well as speaker label. The workflow and corresponding results for the first 23 seconds of one selected call are shown in Figure 1, where the original input is a call audio sample. After data preparation, segments of noise and silence are discarded. This call sample is segmented into 4 utterances. The audio streams are from the original audio and split based on the start and end time of each utterance. Auto transcriptions are more likely to be ungrammatical if the recording quality is bad or the conversation contains words that ASR cannot identify or the speakers do not express themselves clearly. The ungrammatical transcriptions usually occur in customer parts and the frequency of ungrammaticality varies from case to case. Although the sentiment recognition of a whole call tends to be robust with respect to speech recognition errors, the sensitivity of each utterance analysis to ASR errors is not reparable given our study. The speaker labels are from ASD output which can be misclassified because of the occurrence of speakers overlapping or speakers with similar acoustic features. Conversation sentiment pattern study can be misleading due to the misclassified ASD output, although misclassified ASD is rare.

This process allows us to study features from both modalities: transcribed words and acoustics. Distinguishing different parties gives us the ability to study the temporal sentiment transitions of individual speakers and interactions among speakers in a conversation. However, since the data preparation is part of the pipeline described in section 3.4, which runs in real-time, sentiment analysis must rely on error-prone ASR and ASD outputs.

3.3 Sentiment Annotation

Sentiment annotation is a challenging task as the label depends on the annotators' perspective, and the differences inherent in the way people express emotions. The sentiment is opinion-based, not fact-based. This study aims at identifying negative expressions in calls, especially angry customers who are not satisfied with the services, or the business rules, or the systems of rules. By identifying and studying these types of cases, the business can improve call center services and fix the possible business or system issues.

Guidelines are set up for the annotation. The customer negative tag is for negative emotions (e.g. "I hate the system"), attitudes (e.g. "I am not following you"), evaluations (e.g. "your service is the worst"), and negative facts caused by other parties (e.g. "I never received my card"). Other negative facts are not considered as negative (e.g. "My wife died, I need to remove her from my dependents"). The guidelines for CSRs are different. Well trained CSRs usually do not respond negatively, but there are cases that they cannot help the customers. We identify these cases as negative. Cases where a CSR cannot help the customer usually involve business process or system issues.

The sentiment is not always explicit in the text. Borderline linguistic utterances stated loudly and quickly are usually identified as negative (e.g. the utterance "Trust me, it could be done" is classified as negative, since it is in the context that the representative fails to help the customer to enroll in the health plan, and in the audio, the customer is irritated). In all the multimodal sentiment analysis, the labels of all modalities are kept consistent for the same utterance. In our data annotation process, we also keep both text and audio labels that agree with each other and the annotation is based on the audio segments.

3.4 Semi-supervised Learning Annotation Pipeline

To successfully run and train analytical models, massive quantities of stored data are needed. Creating large annotated datasets can be a very time consuming and labor-intensive process. To keep

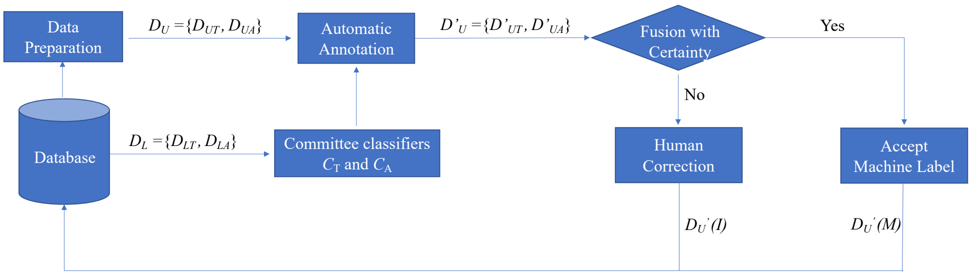


Figure 2: Semi-supervised learning annotation pipeline

the human annotation effort to a minimum, a semi-supervised learning annotation scheme is applied to tag the polarity of utterances as negative, or non-negative. Figure 2 illustrates the process which is similar to active learning annotation. It takes as input a set of labeled examples D_L including text D_{LT} and audio D_{LA}, as well as a larger set of unlabeled examples $D_U = \{D_{UT}, D_{UA}\}$, and produces committee classifiers $C = \{C_T, C_A\}$ and a relatively small set of newly labeled data $D'_U(I)$ and $D'_U(M)$ (Olsson, 2008).

Semi-supervised learning annotation cooperates with humans and machines and combines both semi-supervised learning and multiple classifiers approach for corpus annotation. This pipeline consists of several steps: data generation to obtain D_U (Section 3.2), model training for both modalities to obtain C_T and C_A using D_{LT} and D_{LA} (Section 4), model deployment to get machine label $D'_U = \{D'_{UT}, D'_{UA}\}$, model fusion (Section 5) and results evaluation to decide whether to accept machine label $D'_U(M)$ or ask a human annotator for classifications of the utterances to obtain $D'_U(I)$, then move $D'_U(I)$ and $D'_U(M)$ from D'_U to D_L. It is cyclical and iterative as every step is repeated to continuously improve the accuracy of the classifier and achieve a successful algorithm.

Note, the classifiers in committee $C = \{C_T, C_A\}$ are modified based on D_L in each iteration. The annotation process starts with 20 calls selected from the service center by human domain experts, 20 calls are chunked to 1410 segments via data preparation processing and annotated by three annotators manually as D_L. For the first three iterations, set C_T={Support Vector Machine (SVM), VADER, AWSSA[*], AWSCC[†], GoogleSA[‡]} requires a small size of training data or no extra training data. As the size of D_{LT} increases, we form a new committee C_T = {SVM, Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BLSTM)}. These classifiers are described in Section 4.1. Section 4.2 introduced C_A = {Elastic-Net Regularized Generalized Linear Models (Elastic-Net), K-Nearest Neighbors (KNN), Random Forest (RF), Gaussian Mixture Model (unsupervised GMM) }. In the later iterations, Recurrent Neural Networks (RNN) such as LSTM and BLSTM are applied.

If one call has a long duration ($T > 10$ minutes) and a high percentage of negative utterances based on D'_U ($> 40\%$ for customer or $> 20\%$ for CSR), then we say this call is potentially negative and informative. We then ask an annotator to manually correct the annotated tags D'_U by listening to the call, and move the results $D'_U(I)$ to D_L. For all the other calls, we only keep the utterances where classifiers all agree as $D'_U(M)$. We then remove chunks that are too short (duration $<$ 1s) or too long (duration >20s). Finally, we discard chunks where the annotator cannot discern classification.

Using the pipeline, 6,565 negative and 10,322 non-negative call clips are annotated as the training dataset. The training data D_{LT} still include transcription errors, even though the threshold discussed in the above paragraph is set to eliminate these utterances to add to the training dataset. In addition, 18,705 cleaned text chat data collected from chat windows are also added to D_{LT} via the annotation pipeline to improve the C_T accuracy. Instead of checking fusion with certainty, we only keep the utterances with classifiers in C_T all agree as $D'_U(M) = D'_{UT}(M)$.

Because of the quality of the calls, the poor performance of the ASR for some cases, and the threshold used to annotate the utterances, more than half of the original call segments are discarded[*], and 18,705 text chat data are added to

[*] AWS Comprehend Sentiment Analysis API

[†] AWS Custom Classification API

[‡] Google Language Sentiment Analysis API

[*] The accuracy on the test data decreases by 8% when including all the call segments in the training dataset.

D_{LT}={transcription data, *chat data*} without the corresponding audio files in D_{LA}. It is hard to consider the context of the conversation since the segments are not continuing in the training dataset. Therefore, conversation models are not considered in our committee classifiers C.

4 Bimodal Sentiment Analysis

To model information for sentiment analysis from calls, we first obtain the streams corresponding to each modality via the methods described in Section 3.2, followed by the extraction of a representative set of features for each modality. These features are then used as cues to build classifiers of binary sentiment.

4.1 Sentiment Analysis of Textual Data

General approaches such as sentiment lexicons and sentiment APIs are easy to apply. Both approaches are employed in C_T to monitor the utterance prediction labels in the early stage of semi-supervised learning annotation to extend training data.

VADER (Hutto and Gilbert, 2015) is a simple rule-based model for general sentiment analysis. The results have four categories: compound, negative, neutral, and positive. We classify utterances with negative output as negative, neutral and positive as nonnegative[†] so that it is consistent with BSCD annotation. This model has many advantages, such as being less computationally expensive and easily interpretable. However, one of the main issues with only using lexicons is that most utterances do not contain polarized words. The utterances without polarized words are usually classified as neutral or nonnegative[‡].

Sentiment analysis API is another way to classify sentiment without extra training data. Amazon offers Sentiment Analysis in Amazon Comprehend (AWSSA), which uses machine learning to find insights and relationships in a text. The result returns Mixed, Negative, Neutral, or Positive classification. To be consistent with the BSCD we created, Neutral and Positive are combined as one class: nonnegative[†]. Another sentiment analysis on Google Cloud Natural Language API (GoogleSA) also performs sentiment analysis on text. Sentiment analysis attempts to determine the overall attitude and is represented by numerical scores and magnitude values. We simply set utterances with negative scores as negative and nonnegative otherwise.

For machine learning-oriented techniques by linguistic features, we evaluated well-known SVM, LSTM, and BLSTM models. Since the data is unbalanced and we want the model to focus more on the negative class, we apply weighted loss functions during the training. Hyperparameters are tuned for each model, and ensemble models are also developed by taking the weighted majority vote.

4.2 Sentiment Analysis of Acoustic Data

Feature engineering heavily relies on expert knowledge about data features. To better understand the human hearing process, we study the acoustic features based on human perception. Three perceptual categories are described in this section. Their corresponding features are usually short-term based features that are extracted from every short-term window (or frame). Long-term features can be generated by aggregating the short-term features extracted from several consecutive frames within a time window. For each short-term acoustic feature, we calculated nine statistical aggregations: mean, standard deviation, quantiles (5%, 25%, 50%, 75%, 95%), range (95%-5% quantile), and interquartile range (75%-25% quantile) to get the long-term features of each segment.

- **Loudness** is the subjective perception of sound pressure which is related to sound intensity. Amplitude and mean frequency spectrum features are extracted to measure loudness. The greater the amplitude of the vibrations, the greater the amount of energy carried by the wave, and the more intense the sound will be.
- **Sharpness** is a measure of the high-frequency content of a sound, the greater the proportion of high frequencies the sharper the sound. Fundamental frequency (pitch) and dominant frequency are extracted.
- **Speaking rate** is normally defined as the number of words spoken per minute. In general, the speaking rate is characterized by different parameters of speech such as pause and vowel durations. In our study, speaking rate is measured by pause duration, character per second (CPS), and word per second (WPS) which are calculated as following

[†] Utterances with compound or mixed class are very few, and they are discarded to keep the training data clear.

[‡] This conclusion is verified by the high Rec(+) and low Rec(-) shown in table 1.

for the ith segment:

$$\text{Pause duration}_i = \frac{T_i^{silence}}{T_i^{total}}$$

$$\text{CPS}_i = \frac{N_i^{character}}{T_i^{total}}, \qquad \text{WPS}_i = \frac{N_i^{word}}{T_i^{total}}$$

where for segment i, T_i denotes the time, and N_i denotes the number of characters or words in the corresponding transcription. Pause duration can be interpreted as the percentage of the time where the speaker is silent in each segment. The three variables are aggregated statistics, long-term features.

In nonnegative cases, speakers are in a relaxed and normal emotional state. An agitated or angry emotional state speaker is typically characterized by increased vocal loudness, sharpness, and speaking rate. C_A ={Elastic-Net, KNN, RF, GMM} are built based on the 39 selected features.

Hand-crafted features are generally very successful for specificity sound analysis tasks. One of the main drawbacks of feature engineering is that it relies on transformations that are defined beforehand and ignore some particularities of the signals observed at runtime such as recording conditions and recording devices. A more common approach is to select and adapt features initially introduced for other tasks. A now well-established example of this trend is the popularity of MFCC features (Serizel et al., 2018). In our experiments, MFCC is extracted from each segment and fed to RNN models in later iterations with $|D_{LA}| > 10,000$.

5 Fusion

There are two main fusion techniques: feature-level fusion and decision-level fusion. In our experiments, we employ decision-level fusion. Decision-level fusion has many advantages (Poria et al., 2015). One benefit of the decision-level fusion is we can use classifiers for text and audio features separately. The unimodal classifier can use data from another communication channel of the same type to improve its accuracy, e.g. text data from the chat windows is borrowed to improve the C_T accuracy in our study. Separating modalities permit us to use any learner suitable for the particular problem at hand. In practice, the two unimodal classifiers can be applied separately, e.g. to analyze text data from chat windows $D_U = D_{UT}$, apply C_T only to get sentiment labels D'_{UT}, then add $D'_{UT}(M)$ to D_{LT}. Another benefit of the decision-level fusion is its processing speed since fewer features are used for each classifier and separate classifiers can be run in parallel.

Decision-level fusion usually adds probabilities or summarized predictions from each unimodal classifier with weights or takes the majority voting among the predicted class labels by unimodal classifiers.

In this paper, various fusion methods are evaluated, including two novel approaches that use linguistic ensemble results as the baseline, while then checking acoustic results to modify classification decisions. In Fus1, if the audio ensemble classifies negative and one or more text models classifies negative, we then reclassify the result to negative. In Fus2, if the audio ensemble classifies a sample as negative, we then reclassify the result to negative directly without checking the linguistic modality. The Fus1 and Fus2 approaches are proposed, because for borderline linguistic utterances, acoustic features are more important than linguistic features to understand the spoken intention of the speaker.

6 Experiment Results

The test dataset consists of 21 calls with 1,890 utterances, which are manually annotated for negative (848) and nonnegative (1,042).

6.1 Evaluation Measures

As evaluation measures, we rely on accuracy and weighted F1-score, which is the weighted harmonic mean of precision and recall. Precision is the probability of returning values that are correct. Recall, also known as sensitivity is the probability of relevant values that the algorithm outputs.

As shown in Table 1, general approaches in C_T, Vader and APIs, tend to have a low negative recall. The semantic knowledge based classifiers have more than 20% higher weighted F1-score than the general approaches. The classifiers are trained on D_{LT}={transcription data, chat data}. The overall weighted F1-score is more than 10% higher than the classifiers trained on call transcription only data[§].

BLSTM on MFCC performs better than C_A = {Elastic-Net (penalty $0.2||\beta||_1 + 0.4||\beta||_2^2$), KNN ($k = 3$), RF, GMM} on acoustic features. Using audio features alone, a weighted F1-score of 0.584

[§] Weighted F1-scores are 0.718 (SVM), 0.719 (LSTM) and 0.714 (BLSTM).

Methods	Text						Audio				
	SVM	LSTM	BLSTM	Vader	AWSSA	GoogleSA	Elastic-Net	KNN	RF	GMM	BLSTM
Acc.	0.814	0.853	0.843	0.498	0.651	0.637	0.570	0.544	0.585	0.546	0.601
F1 (w)	0.814	0.852	0.842	0.347	0.628	0.615	0.500	0.534	0.549	0.500	0.584
Prec.(+)	0.770	0.802	0.781	0.494	0.594	0.586	0.528	0.518	0.541	0.513	0.561
Prec.(-)	0.871	0.92	0.934	0.742	0.821	0.779	0.860	0.589	0.741	0.685	0.693
Rec. (+)	0.886	0.929	0.946	0.991	0.908	0.881	0.964	0.697	0.883	0.872	0.811
Rec. (-)	0.746	0.779	0.745	0.024	0.404	0.402	0.205	0.402	0.309	0.252	0.402

Table 1: Binary classification of sentiment polarity on test data: Accuracy (Acc.), weighted F1-score (F1 (w)), precision (Prec.) and recall (Rec.) for the nonnegative (+) and negative (-) classes

Methods	Ensemble			Fusion	
	Text	Audio	T+A	Fus1	Fus2
Acc.	0.851	0.586	0.846	0.858	0.871
F1 (w)	0.851	0.525	0.846	0.858	0.871
Prec.(+)	0.779	0.531	0.800	0.790	0.818
Prec.(-)	0.949	0.927	0.896	0.946	0.933
Rec. (+)	0.953	0.979	0.894	0.950	0.933
Rec. (-)	0.761	0.240	0.804	0.777	0.817

Table 2: Binary classification of sentiment polarity on both linguistic and acoustic modalities

can be reached, which is acceptable considering that the real world audio-only system exclusively analyzes the tone of the speaker's voice and doesn't consider any language information.

The acoustic modality is significantly weaker than the linguistic modality. Usually, speakers' tones are not signifcantly different from the tones under normal emotional state even the content is negative (e.g. "We messed up." with negative tag). 97% of the segments with correct D'_{UT} but wrong D'_{UA} have negative as true tag. The other 3% are the nonnegate segments with emphasized words (e.g. " But I do have a newborn coming." with nonnegtive tag).

In most cases, text already includes enough information to judge the sentiment. A few observed typical situations leading to linguistic modality misclassification are the presence of misleading linguistic cues caused by overlapping or other issues (e.g. ASR "Customer: I love it. It can be done." and true transcription "CSR: I... Customer: Drop it. It can be done." with negative tag), ambiguous linguistic utterances whose sentiment polarity are highly dependent on the context described in earlier or later part of the call (e.g. "But I got a call from your service center today apologizing, saying, Yeah, we made a mistake." with nonnegative tag), or nonnegative linguistic utterances stated angrily (e.g. "So I think you should honor those amounts." with negative tag).

In order to achieve better accuracy, we combine the two modalities together to exploit complementary information. We simply combine results of the three semantic knowledge based classifiers and all the five audio classifiers by taking the weighted majority vote. The T+A ensemble results are shown in Table 2 and they do not improve when compared to the unimodal text ensemble results.

Since the unimodal performance of linguistic modality is notably better than acoustic modality, our decision-level fusion methods use linguistic ensemble results as the base-line, while acoustic results are used as supplemental information to calibrate each classification. Fus1 reclassifies the ambiguous linguistic utterances, and Fus2 reclassifies the nonnegative/ambiguous linguistic utterances based on audio ensemble classifies. The two novel fusion approaches discussed in Section 5 are tested. The Fus2 bimodal system yields a 2% improvement in weighted F1-score than the text unimodal system.

McNemar's test is applied to compare the accuracy of text only results D'_{UT} and Fus2 results D'_{UF2}

$$\chi^2 = \frac{(14-52)^2}{14+52} = 21.88,$$

where the number of segments with correct D'_{UT} wrong D'_{UF2} is 14, and wrong D'_{UT} correct D'_{UF2} is 52. The McNemar's test gives $\chi^2 = 21.88$ and $P < 0.001$, which implies a statistically significant effect by adding acoustic features using the Fus2 approach.

The acoustic modality provides important cues to identify borderline linguistic segments with negative emotions. Our results show that relying on the joint use of linguistic and acoustic modalities allows us to better sense the sentiment being expressed as compared to the use of only one modality at a time. The acoustic feature analysis helps us to better understand the spoken intention of the

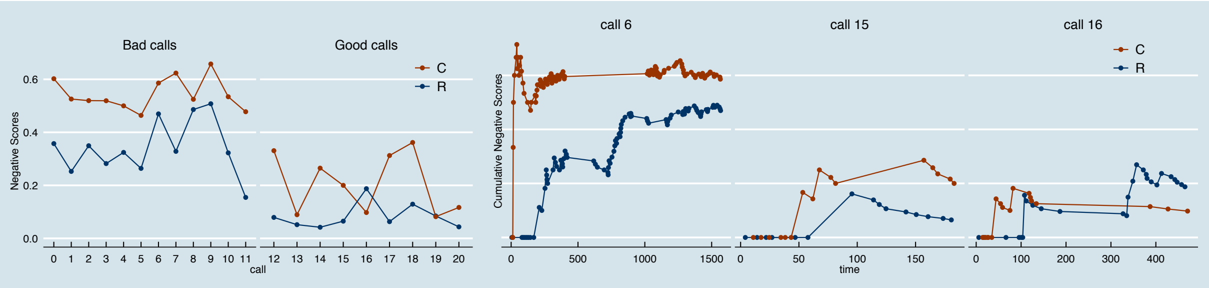


Figure 3: The (cumulative) negative score pattern between customers (C) and CSRs (R)

speaker, which is not normally expressed through text.

6.2 Tempo Sentiment Pattern

The sentiment is not only regarded as an internal psychological phenomena but also interpreted and processed communicatively through social interactions. Conversations exemplify such a scenario where inter-personal sentiment influences persist.

The left panel in Figure 3 shows the negative scores of customers and CSRs in 21 test calls. The negative score, a weighted negative segment percentage, is calculated to analyze the overall sentiment. Weights 0.8, 1, and 1.2 are assigned to the first third, second third and last third of each call. Since long pauses in calls are discarded in the data preparation process, these segments do not have sentiment labels and do not contribute to the negative score. The negative scores of CRSs are commonly lower than customers', and usually high negative scores for customers correspond to high negative scores for CSRs. We can conclude from the figure that sentiment can be affected by other parties during a conversation.

To further analyze the interactions between customers and CSRs, the cumulative negative scores for call 6, 15, and 16 are drawn on the right panel of Figure 3. The x-axis shows time of the whole call in seconds including noise and long pauses. Call 6 shows the sentiment patterns of a typical bad call, which is characterized by long duration and long pauses. The two long pauses are from 444s to 607s and from 921s to 1008s. Between the two long pauses, there are three customer and CSR overlapping segments, but the Automatic Speaker Diarization recognizes all of them as CSRs. The customer has a high negative score from beginning to end, and the CSR fails to help the customer during the call. Call 15 is a typical good call. The overall negative score is low and the negative score pattern goes down for both the customer and the CSR, which means the problem is resolved by the end of the call. Call 16 is another type of call, in which the customer does not get angry even though the CSR is unable to solve his/her issues.

7 Discussion and Future Work

A new dataset BSCD consisting of real-world conversation, the service calls, is introduced. Human communication is a dynamic process, and our eventual goal is to develop a bimodal sentiment analysis engine with the ability to learn the temporal interaction sentiment patterns among conversation parties. In the process of fusion, we have approached the study of audio sentiment analysis from an angle that is somewhat different from most people's.

Future research will concentrate on evaluations using larger data sets, exploring more acoustic feature relevance analysis, and striving to improve the decision-level fusion process.

A call is constituent of a group of utterances that have contextual dependencies among them. However, in our semi-supervised learning annotation pipeline, about half of the segments in calls are discarded. Therefore the interdependent modeling is out of the scope of this paper and we include it as future work.

Acknowledgements

The author wishes to express sincere appreciation to Sony SungChu for the support of the project and the comments that greatly improved the manuscript, and the anonymous reviewers for their insightful suggestions to help clarify the manuscript.

References

Gustavo Aguilar, Viktor Rozgic, Weiran Wang, and Chao Wang. 2019. Multimodal and multi-view models for emotion recognition. arXiv:1906.10198. Version 1.

Samah Alhazmi, Bill Black, and J McNaught. 2013. Arabic SentiWordNet in relation to SentiWordNet 3.0. *International Journal of Computational Linguistics*, 4:1–11.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Erik Cambria, J. Fu, Federica Bisio, and Soujanya Poria. 2015. AffectiveSpace 2: Enabling affective intuition for concept-level sentiment analysis. *Proc. AAAI*, pages 508–514.

Erik Cambria, Devamanyu Hazarika, Soujanya Poria, Amir Hussain, and Rbv Subramanyam. 2017. Benchmarking multimodal sentiment analysis. arXiv:1707.09538. Version 1.

Erik Cambria, Robyn Speer, C. Havasi, and Amir Hussain. 2010. SenticNet: A publicly available semantic resource for opinion mining. *AAAI Fall Symposium - Technical Report*, pages 14–18.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI meeting corpus: A pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg. Springer-Verlag.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. openSMILE – the munich versatile and fast open-source audio feature extractor. *MM'10 - Proceedings of the ACM Multimedia 2010 International Conference*, pages 1459–1462.

Florian Eyben, Martin Wllmer, and Björn Schuller. 2009. openEAR - introducing the munich open-source emotion and affect recognition toolkit. *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference*, pages 1 – 6.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report*, 150.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: Interactive conversational memory network for multimodal emotion detection. pages 2594–2604.

C.J. Hutto and Eric Gilbert. 2015. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*.

Manas Jain, Shruthi Narayan, Pratibha Balaji, Abhijit Bhowmick, and Rajesh Muthu. 2018. Speech emotion recognition using support vector machine.

Margarita Kotti and Fabio Paternò. 2012. Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema. *Int J Speech Technol*, 15.

Bing Liu. 2010. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing, Second Edition*.

Bing Liu. 2012. Sentiment analysis and opinion mining. In *Synthesis Lectures on Human Language Technologies*.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An attentive RNN for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:6818–6825.

Daniel McEnnis, Cory McKay, Ichiro Fujinaga, and Philippe Depalle. 2005. jAudio: An feature extraction library. *Proceedings of the International Conference on Music Information Retrieval*, pages 600–603.

Brian McFee, Colin Raffel, Dawen Liang, Daniel Ellis, Matt Mcvicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. *in Proceedings of the 14th Python in Science Conference*, 8:18–24.

Gary Mckeown, Michel Valstar, Roddy Cowie, and Maja Pantic. 2010. The SEMAINE corpus of emotionally coloured character interactions. *2010 IEEE International Conference on Multimedia and Expo, ICME 2010*, pages 1079–1084.

Gary Mckeown, Michel Valstar, Roddy Cowie, Maja Pantic, and M. Schroder. 2013. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3:5–17.

Finn Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *CoRR*.

Fredrik Olsson. 2008. Bootstrapping named entity annotation by means of active machine learning: a method for creating corpora.

Y Pan, P Shen, and L Shen. 2012. Speech emotion recognition using support vector machine. *Int. J. Smart Home*, 6:101–108.

Bo Pang and Lillian Lee. 2004a. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Computing Research Repository - CORR*, 271-278:271–278.

Bo Pang and Lillian Lee. 2004b. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain.

James Pennebaker, Cindy Chung, Molly Ireland, Amy Gonzales, and Roger Booth. 2007. The development and psychometric properties of LIWC2007.

James Pennebaker, M. Francis, and R. Booth. 2001. Linguistic inquiry and word count (LIWC): LIWC2001. 71.

Soujanya Poria. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2539–2544, Lisbon, Portugal. Association for Computational Linguistics.

Soujanya Poria, Erik Cambria, Amir Hussain, and Guang-Bin Huang. 2014a. Towards an intelligent framework for multimodal affective data analysis. *Neural Networks*, 63.

Soujanya Poria, Alexander Gelbukh, Erik Cambria, Amir Hussain, and Guang-Bin Huang. 2014b. EmoSenticSpace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems*, 69.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. arXiv:1905.02947. Version 1.

Cícero dos Santos and Maíra Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. 2012. AVEC 2012 – the continuous audio/visual emotion challenge. *ICMI'12 - Proceedings of the ACM International Conference on Multimodal Interaction*.

Romain Serizel, Victor Bisot, Slim Essid, and Gaël Richard. 2018. Acoustic features for environmental sound analysis. *Computational Analysis of Sound Scenes and Events*, pages 71–101.

Caifeng Shan, Shaogang Gong, and Peter W. Mcowan. 2007. Beyond facial expressions: Learning human emotion from body gestures. In *in Proc. British Machine Vision Conf.*

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew .Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. *EMNLP*, 1631:1631–1642.

Philip Stone, Dexter Dunphy, Marshall Smith, and Daniel Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*, volume 4.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. *Proceedings of the 4th International Workshop on the Semantic Evaluations (SemEval 2007).*

Jerome Sueur, Thierry Aubin, and Caroline Simonis. 2008. Seewave: a free modular tool for sound analysis and synthesis. *Bioacoustics*, 18:213–226.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation (formerly Computers and the Humanities)*, 39:164–210.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of HLT/EMNLP*.

Z. Zeng, J. Tu, M. Liu, T. S. Huang, B. Pianfetti, D. Roth, and S. Levinson. 2007. Audio-visual affect recognition. *IEEE Transactions on Multimedia*, 9(2):424–428.