Combining Two Perspectives on Classifying Multimodal Data for Recognizing Speaker Traits

Moitreya Chatterjee
CrowdShipping Inc.
Los Angeles, CA, USA
metro.smiles@gmail.com

Sughyun Park
USC Institute for Creative Technologies
Playa Vista, CA, USA
park@ict.usc.edu

Louis-Philippe Morency
LTI, Carnegie Mellon University
Pittsburgh, PA, USA
morency@cs.cmu.edu

Stefan Scherer
USC Institute for Creative Technologies
Playa Vista, CA, USA
scherer@ict.usc.edu

ABSTRACT

Human communication involves conveying messages both through verbal and non-verbal channels (facial expression, gestures, prosody, etc.). Nonetheless, the task of learning these patterns for a computer by combining cues from multiple modalities is challenging because it requires effective representation of the signals and also taking into consideration the complex interactions between them. From the machine learning perspective this presents a two-fold challenge: a) Modeling the intermodal variations and dependencies; b) Representing the data using an apt number of features, such that the necessary patterns are captured but at the same time allaying concerns such as over-fitting. In this work we attempt to address these aspects of multimodal recognition, in the context of recognizing two essential speaker traits, namely passion and credibility of online movie reviewers. We propose a novel ensemble classification approach that combines two different perspectives on classifying multimodal data. Each of these perspectives attempts to independently address the two-fold challenge. In the first, we combine the features from multiple modalities but assume inter-modality conditional independence. In the other one, we explicitly capture the correlation between the modalities but in a space of few dimensions and explore a novel clustering based kernel similarity approach for recognition. Additionally, this work investigates a recent technique for encoding text data that captures semantic similarity of verbal content and preserves word-ordering. The experimental results on a recent public dataset shows significant improvement of our approach over multiple baselines. Finally, we also analyze the most discriminative elements of a speaker's non-verbal behavior that contribute to his/her perceived credibility/passionateness.

ICMI '15, November 09-13, 2015, Seattle, WA, USA.

http://dx.doi.org/10.1145/2818346.2820747.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: User/Machine Systems—*Human information processing*; I.5.3 [**Pattern Recognition**]: Clustering—*Algorithms, Similarity Measures*

Keywords

Generative Model; Discriminative Model; Clustering; Kernels; Ensemble; Multimodal

1. INTRODUCTION

Human communication/interaction is a complex process that involve both verbal and non-verbal cues (eye gaze, head gestures, voice pitch, etc.). Very often the motive behind initiating such a communication or interaction is to persuade others' of a person's opinion on a certain topic. Consider a public speech or an online video presenting a person's viewpoint, for instance. In such scenarios, the key to effective communication is, in Aristotle's words, the three components of *Ethos, Pathos* and *Logos* [13]. The *Ethos* of the speaker refers to his/her perceived credibility by the audience, *Pathos* emphasizes the importance of building an emotional connection with the audience by being passionate. Finally *Logos* underscores the importance of presenting a logically cogent argument. While *Logos* often entails an objective presentation of facts, the other two however, can be achieved by the effective use of both verbal and non-verbal communication channels [21].

The proliferation of video-sharing websites (YouTube, Dailymotion, etc.) has meant that more and more of human communication is taking place online. People now share their opinions of almost everything from politics to commercial products through online videos, making it more and more important to study human communication in the online domain. In this work, we therefore focus on recognizing two important speaker traits, namely the degree of *passion* and *credibility*, of a speaker in online reviews of movies. This task raises three important research questions.

The first one pertains to the machine learning component of this task. From the perspective of machine learning, this task presents a two-fold challenge. On the one hand, we need to model the variations and dependencies between the

modalities (text, audio, video), since a simplistic representation of the data without capturing the inter-modality variations, would most likely fail to uncover important latent patterns [23]. On the other hand, we need to ensure that the data is represented using an apt number of effective features. A large number of features for modeling the several complex patterns in the data is both hard to design and also faces the challenge of potential over-fitting [5]. These issues raise the central research question, **Q1**: *Is there an effective technique of representing and classifying multimodal data so that the important inter-modal dependencies are captured?*

In this work, we attempt to address this challenge by proposing a novel ensemble-based classification scheme that combines discriminatory information from two different perspectives on classifying multimodal data. In the first one, we combine the features from multiple modalities in a combined space but assume conditional independence across the modalities. This avoids the task of having to learn too many parameters. In the other one, we project the data to a space of few dimensions using Multiview Canonical Correlation Analysis(MVCCA) [14], such that the correlation between the modalities is explicitly captured. We then classify using a novel non-parametric approach that relies on clustering-based kernel similarity. Our inclination towards a non-parametric approach, stems from the strong theoretical guarantees of performance that they have [9]. Figure 1 presents an overview of our approach.

The second research question pertains to effective representation of text. The conventional text representation approach for analyzing online movie reviews, Bag-of-Words [22, 10], suffers from a lack of semantic-awareness of the content and loss of word ordering [16]. Thus our concern is, **Q2**: *What constitutes an effective feature representation of text data in the context of online reviews?* As a potential solution, we explore a recent style of embedding text features, that has yet remained unexplored in the context of multimodal recognition. This technique maps semantically similar paragraphs as less-distant points in a feature space and also preserves word-ordering [16].

On the side of non-verbal behavior(acoustic, visual cues), we are faced with our third concern. By its very nature such behavior is subtle. We are therefore interested in knowing, **Q3**: *What non-verbal behavioral components allows us to discriminate between passionate/credible reviewers and those who are not?* We attempt to address this by performing an analysis of the descriptors of non-verbal behavior.

The core contribution of this paper is our attempt to answer these 3 research questions.

Our experiments on the recently released Persuasive Opinion Multimedia Corpus(POM) [22] show that our proposed solutions to all three research questions(Q1-Q3) hold promise, for the intended recognition tasks.

2. RELATED WORK

A feature-level stacking technique, called *early fusion* has widely been explored for multimodal recognition [4, 22, 24, 6]. While this technique does combine cues from multiple modalities, however, it does not explicitly model the inter-modality correlations. Furthermore a fusion of features from multiple modalities via *early fusion*, often leads to a blow-up of the number of feature dimensions. This makes it prone to over-fitting(high variance), resulting in poor generalization [5]. Many of these previous approaches resort to either dimension reduction or feature selection techniques to address this issue [24, 22]. On the contrary, we explore a novel idea of an ensemble of two classification approaches, each of which attempts to mitigate this concern in a different way. The first of these incorporates the inter-modality dependencies and variances through a Multiview Canonical Correlation Analysis(MVCCA) [14]-based technique and projects the data to a low-dimensional space, allaying concerns related to dimensionality. While the second classification technique, stacks up the features in a combined space like *early fusion* but assumes inter-modality conditional independence. This avoids the task of having to learn a large number of validated parameters.

Recently, Song et al. [23] successfully explored a technique for multimodal classification that models explicit correlation between the modalities through Canonical Correlation Analysis(CCA) [12]-based techniques. Their approach however, considers only two modalities for fusion. Our MVCCA-based approach, is an extension of CCA for more than 2 modalities. To the best of our knowledge, this application of MVCCA to multimodal (> 2) recognition is novel.

For the task of multimodal recognition, *mixed norm regularization* has shown promise [26]. In this approach, Zhuang et al. regularize data from different modalities using separate norms, viz. L_1 and L_2 respectively. We,On the other hand, in one of our classification approaches, model features from dissimilar modalities using entirely different probability distributions.

Bag-of-Words(BoW) has been a long-standing approach for capturing patterns in text for a diverse range of applications [18, 17, 10, 22]. However, an inherent weakness of this technique is that it only considers frequency of occurrence of the words and does not capture the semantic similarity between them. This might potentially result in the model being perplexed about the topic of the document if prepositions, articles, verbs, etc. occur frequently. Further, BoW also fails to preserve word ordering. We therefore investigate a recent technique for text encoding which maps the text document to a word-order preserving feature space, where semantically similar paragraphs occur as less-distant points [16].

Finally in the context of online movie reviews, Park et al. have probed into "persuasiveness" of the reviewer [22]. We on the other hand, explore "credibility" and "passion" which are two well-known attributes of an eloquent speaker [13]. The reader is referred to Mohammadi et al. for a treatise on how these attributes relate to personality traits [19].

3. DATASET

The Persuasive Opinion Multimedia(POM) Corpus is a recently introduced public dataset of 1,000 online movie review videos crawled from *ExpoTV.com* [22]. Each video consists of a speaker, presenting his/her review of a movie, facing the camera. The crawled videos are controlled for their video and audio qualities, in order to avoid noisy samples. The mean length of the videos in the corpus is 94.38 seconds. This corpus consists of just those reviews where the speaker either assigned 5-stars (most recommended) or 1 or 2-stars (least recommended) for the movie they are reviewing and these direct ratings by the speakers are made available. For 500 of these reviews, the reviewers have assigned 5-stars for the corresponding movie while the remaining 500 have been assigned 1 or 2-stars. The dataset along

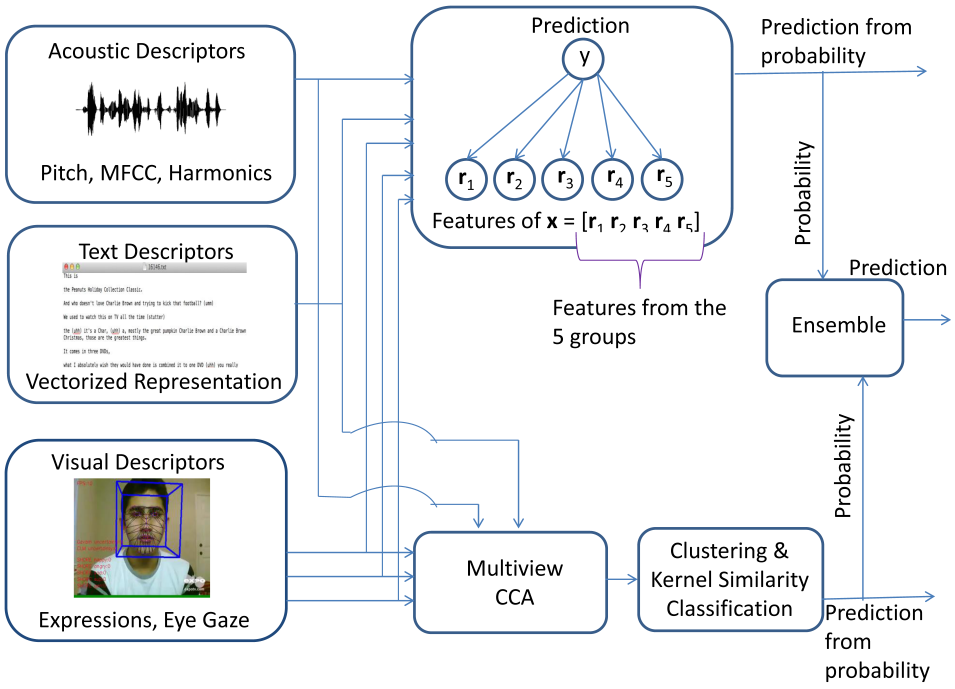

Figure 1: **The overview of our approach. We extract 5 groups of features from audio(1), text(1) and visual(3) channels. These features are then fed into two separate classification pipelines. Finally, an ensemble combines the outputs.**

with the transcriptions of what the reviewers say (including para-verbal markers such as stuttering, filled pauses, etc.) are also made available.

Each review has been annotated for various high-level subjective attributes such as "Persuasion", "Passion", "Credibility", "Confidence", etc. on a Likert-scale of 1(low) to 7(high) by 3 annotators, from Amazon Mechanical Turk. This work being targeted at exploring the qualities of an eloquent speaker, we focus exclusively on the "Passion" and "Credibility" attributes. The inter-coder agreement, as measured by Krippendorff's alpha, is 0.69 and 0.75 for "credibility" and "passion" respectively. This is suggestive of a high inter-coder agreement.

4. PROPOSED APPROACH

4.1 Computational Descriptors

We designed descriptors for the acoustic, visual and verbal modalities. They are as follows:

4.1.1 Descriptors for the Acoustic Modality

Park et al. have successfully explored a set of acoustic features for determining the degree of persuasiveness of a speaker in the context of online movie reviews [22]. We recomputed those features over the full length of the acoustic signals. It involved computing several statistical functionals for the first 5 Formants, the first 24 Mel Cepstral Frequency Coefficients (MFCCs), pitch (F0), and several voice quality measures including Normalized Amplitude Quotient (NAQ), Parabolic Spectral Parameter (PSP), Maxima Dispersion Quotient (MDQ), Quasi-Open Quotient (QOQ), difference between the first two harmonics (H1 - H2), and peak-slope. We used the mean, median, percentiles (10th, 25th, 75th, and 90th), ranges (between min and max, 10th and 90th percentiles, and 25th and 75th percentiles), and standard deviation as the statistical functionals.

4.1.2 Descriptors for the Visual Modality

On similar lines with the acoustic modality, we computed the same set of statistical functionals for automatically tracked facial features. The statistical functionals were computed for discrete emotions (anger, contempt, disgust, fear, joy, sadness, and surprise), valence (positive and negative), several Facial Action Units, eye gaze movements (up/down, left/right), and head movements (about the X, Y and Z axes). A detailed description is available in Park et al. [22]. Different from Park et al. however, we treat the head motion, eye gaze and the rest of the descriptors as three separate groups of features, i.e. we treat them as though they were descriptors from 3 separate modalities. This is because they are completely heterogeneous signals with very different amplitude variations and are computed using three different trackers, namely GAVAM[20], OKAO[3] and FACET[1].

4.1.3 Descriptors for the Verbal Modality

Modeling text using a bag-of-words (BoW) technique has been a long standing approach [18, 17]. Recently, this has also been explored in the context of assessing persuasiveness of online movie reviewers [10]. However this technique does not preserve word-ordering nor does it capture the semantic similarity among words. It treats every word independently and builds a histogram based on *term frequency*. This often results in "significant" words in the context of a certain problem being dominated by prepositions, articles or other words that tend to occur more frequently in any text document. Filtering out a pre-determined list of words or using

the frequency count to trim down the vocabulary are not generalizable from one application to another.

To mitigate these weaknesses, we project the text data into a fixed 100-dimensional continuous feature space where semantically similar paragraphs across documents map as less-distant points [16, 2]. This framework is completely unsupervised and encodes every paragraph as a point in a multi-dimensional feature space. The approach works by comparing the word-ordering in a paragraph to their likelihood of occurrence in a training corpus. This likelihood is used to extract features using a Neural Language Model [7]. Finally, a linear regression like approach maps these features to a vector.

We thus have a total of five(5) sets of features, which we call *feature groups*, audio - 1, text - 1 and visual - 3.

4.2 Classification Technique

Our classification technique relies on an ensemble of two classification approaches that adopt two different perspectives for classifying multimodal data. For the subsequent explanations, we adopt the following notations. We have a set of N training samples, $\mathcal{X} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, where every sample $\mathbf{x_i}$ is represented by a concatenation of the set of features obtained from the five feature groups, $\mathbf{x_i} = [\mathbf{r}_{i_1}, \mathbf{r}_{i_2}, ..., \mathbf{r}_{i_5}]$. y_i is the class label and $y_i \in \{+1, -1\}$. For a test sample $\mathbf{x}$, represented by $\mathbf{x} = [\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_5]$, we want to assign a label y, where $y \in \{+1, -1\}$.

4.2.1 Ensemble Classifier

Our proposed ensemble classifier is a linear weighted combination of two classifiers, called the Modality-Independent Bayesian Classifier and the Similarity Classifier. These two classifiers adopt two different perspectives on classifying multimodal data. The ensemble allows us to combine the discriminative powers of both the models. For the task of classification, we compute the posterior probability for $\mathbf{x}$ as:

$$P_E = P(y = 1|\mathbf{x}) = \rho P_1 + (1 - \rho)P_2; \rho \in [0, 1] \qquad (1)$$

where P_1 and P_2 are the posterior probabilities for $\mathbf{x}$ obtained from the Modality-Independent Bayesian Classifier and the Similarity Classifier, respectively. These two classifiers are described in subsequent sections. We then, compare P_E to $P_E^c = P(y = -1|\mathbf{x}) = 1 - P_E = \rho(1 - P_1) + (1 - \rho)(1 - P_2)$ and determine which one is greater and assign the label to the test sample accordingly. ρ is chosen to maximizes the classification accuracy on the training data.

4.2.2 Modality-Independent Bayesian Classifier

Our first perspective on the data is that of conditional independence of the modalities(feature groups). This assumption is directed at addressing the issue of potential over-fitting of a large number of model parameters. Such scenarios are commonplace in early fusion settings, which attempt to model a full association among all the features across all modalities [5].

Mixed norm regularization is a classification approach that works by combining decisions from multiple modalities. For the task of learning the decision boundary, the parameters for each modality are regularized with a different norm, viz. L_1 and L_2 [26]. This allows the model to capture the variations within different modalities, to some extent. Inspired by such approaches, we model features from different feature groups using entirely different distributions. The posterior probability for $\mathbf{x}$ is computed as,

$$P(y|\mathbf{x}) \propto P(\mathbf{x}|y)P(y) = \prod_{i=1}^{5} P(\mathbf{r}_i|y)P(y),$$

This is a generative technique of modeling and its use is inspired by recent successes of such approaches for various learning tasks [15]. Our approach might reminisce the reader about the popular Naive Bayes classifier [9]. However different from that, we do not assume full conditional independence among all the features within a feature group. In fact, we represent each of the class-conditionals for every feature group, $P(\mathbf{r}_i|y)$ by a separate multidimensional Gaussian Distribution, $\mathcal{N}(\mathbf{r}_i; \mu, \Sigma)$, with a fully general covariance matrix Σ, learned from the training data. We then pick the greater of $P_1 = P(y = 1|\mathbf{x})$ and $P_1^c = P(y = -1|\mathbf{x})$ for label assignment. This results in a model that not only combines cues from multiple modalities but also drastically cuts the number of validation parameters. This P_1 is also used for the Ensemble classification by substituting it in Equation 1. We subsequently refer this approach as Cls. 1.

4.2.3 Similarity Classifier

In our second perspective, we explicitly capture the inter-modality correlations in the process of classifying. Our classification technique, is a non-paramteric one and banks on clustering and a kernel similarity ratio. The 3-step procedure for this is as follows:

• **MVCCA**: We first perform Multiview Canonical Correlation Analysis (MVCCA) to project the data to a 5-dimensional space where the 5 features groups are correlated the most [14]. This achieves two objectives: a) Maps the data to a space where the cross-modality features correlate, thereby making them more homogeneous (see Figure 2); b) Performs dimensionality reduction. The regular CCA [12] works for two feature groups by computing projection vectors for each that maximize the correlation between the two. For 3 or more feature groups however, there are several options such as maximizing the sum of correlations, maximizing the variance in the data in the projected space or minimizing the variance, minimizing the sum of squared errors, etc. [14]. Each of these approaches optimizes a different objective function. We choose to maximize the sum of correlations because of the ease of solving the corresponding optimization problem. Let $\mathcal{X} \in \mathbb{R}^{N \times (D_1 + ... + D_5)}$ be the training set of N samples, where $D_1, ..., D_5$ are the dimensions of each feature group. We compute the set of projection vectors $\mathbf{h}_1, ..., \mathbf{h}_5$ where $\mathbf{h}_1 \in \mathbb{R}^{D_1 \times 1}$ and so on s.t.:

$$\arg\max_{\mathbf{h}_1,...,\mathbf{h}_5} \sum_{i \neq j; i,j=1}^{5} \mathbf{h}_i^T \Sigma_{ij} \mathbf{h}_j;$$

$$s.t.\ \mathbf{h}_i^T \Sigma_{ii} \mathbf{h}_i = 1;\ \forall i = 1, ..., 5,$$

where $\Sigma_{ij} \in \mathbb{R}^{D_i \times D_j}$ is the covariance matrix between feature groups i and j. This optimization is solved by the Generalized Eigen Value technique [25].

After having projected the data to a 5-d MVCCA space, we cluster the data:

• **Clustering in the MVCCA Space**: We first cluster the training data in the MVCCA space. We perform clustering using the standard Gaussian Mixture Model (GMM)-Expectation Maximization (EM) technique [8] followed by

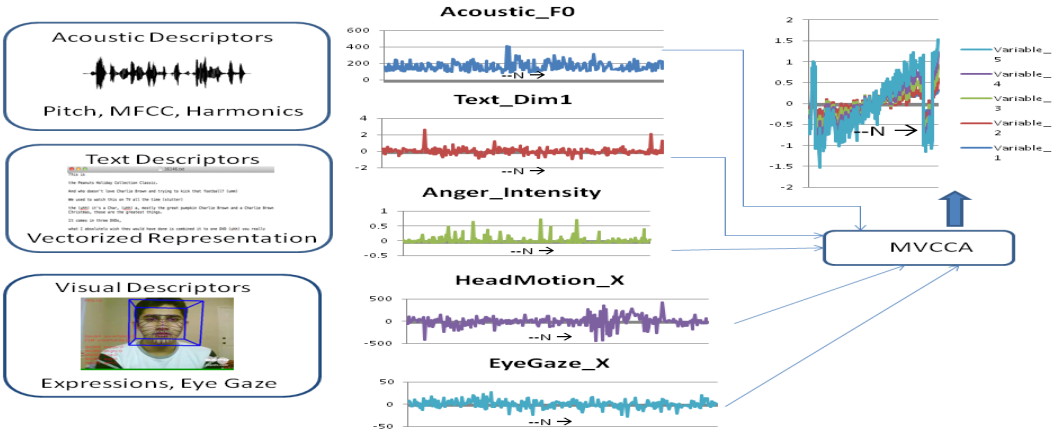

Figure 2: **This figure shows how signals from multiple groups(5 for our case) are projected into a common space where their mutual correlation is the highest. This is for our experiments with "passion."**

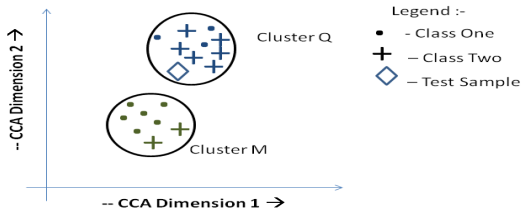

Figure 3: **The test sample is assigned to Cluster Q and thus all samples in Cluster M may be discarded for classification.**

the crisp assignment of every sample to a cluster, i.e. every Gaussian component is considered as one cluster and every sample is assigned to exactly one cluster. The number of mixture components are automatically chosen based on whichever number, in the range of 1 to 20, minimizes the Akaike Information Criterion (AIC) scores.

Now that we have our clusters on the training data in place, we assign each test sample $\mathbf{x}$ (projected in MVCCA space) to one of these clusters based on the most probable candidate. Assuming there are T clusters, this step may be defined as:

$$\arg \max_{i=\{1,\ldots,T\}} P(\mathbf{x}; \mu_i, \Sigma_i),$$

where μ_i, Σ_i are the mean and the covariance matrix of the Gaussian distribution corresponding to the i^{th} cluster.

Our motivation for clustering the data is two-fold: a) Our classification technique for the Similarity Classifier is non-parametric, quite like k-Nearest Neighbor(kNN) [9]. However, unlike kNN, which is constrained to choose a fixed number of neighbors for decision making(k), we dynamically assign all the training samples in the same cluster as the test, to be the test sample's "neighbors". Clustering incorporates this flexibility.; b) Speed-gain in the process of classification.

• **Classification**: For the final classification, we adopt a non-parametric approach. Our reason for being inclined towards such an approach is because theoretically, the empirical risk for non-parametric classifiers is bounded to be within twice that of a hypothetical optimal classifier [9]. This gives an assurance in terms of classification performance of such algorithms. For classifying the test sample, we look into that subset of training samples that are in the same cluster as the test (say Q). The training samples in Cluster Q have the most similar feature appearance as the test and thus we discard all other training samples. This prevents the dissimilar training samples from influencing the inference process and also adds to the computational speed of our algorithm (see Figure 3). Now we compute a normalized similarity score between the test sample and the samples of each of the classes in cluster Q and see to which one it is more similar. In order to capture non-linear similarity patterns, we compute this similarity using kernel mappings. The normalization allows for a probabilistic interpretation of the similarity score. The posterior probability for the test sample $\mathbf{x}$ is computed by using the kernel similarity ratio as follows:

$$P_2 = P(y = 1|\mathbf{x}) = \frac{\sum_{\mathbf{x}_i \in \mathcal{X}, \mathbf{x}_i \in Q, y_i = 1} K_{Gauss}(\mathbf{x}, \mathbf{x_i}; \Sigma_Q)}{\sum_{\mathbf{x}_i \in \mathcal{X}, \mathbf{x}_i \in Q} K_{Gauss}(\mathbf{x}, \mathbf{x_i}; \Sigma_Q)},$$

where

$$K_{Gauss}(\mathbf{x}, \mathbf{x_i}; \Sigma_Q) = \exp(-\mathbf{x}^T (\Sigma_Q + \lambda I)^{-1} \mathbf{x_i}/2),$$

is a Gaussian kernel whose variance (Σ_Q) is the variance of the Gaussian distribution that was fitted to cluster Q in the previous step, I is the identity matrix and λ is the regularization term, typically with a very low value. Label assignment is done by comparing P_2 with $P_2^c = 1 - P_2$ and choosing the higher one. This P_2 is also plugged into Equation 1 for computing the ensemble classifier. We subsequently refer to this approach as Cls. 2.

5. EXPERIMENTS

5.1 Ground-Truth Labels

This work aims at exploring the differences in verbal and non-verbal behavior between highly passionate and credible reviewers and their counterparts on the other side of the spectrum. For our experiments, one for passion and the other for credibility, we therefore choose only those videos that had a mean annotator rating of 5.5 or more(highly passionate/credible) or 2.5 or less(weakly passionate/credible). For our experiments with passion, this resulted in a total of 291 videos(116 passionate and 175 phlegmatic) while for credibility, we ended up with a total of 233 videos(122 of the reviews deemed highly credible and 111 as not).

5.2 Methodology

We performed experiments to address our three research questions (Q1-Q3).

To address Q1, we conduct two separate classification experiments for recognizing "passion" and "credibility". Our experiments for passion and credibility, are based on the ground-truth labels obtained using the procedure described above. We perform *speaker and movie-independent* 5-fold cross-validation experiments, with separate training (4 partitions) and test partition (1 partition) for each. Thus we perform, 5 separate classification experiments for each of passion and credibility and report the mean classification accuracy. Our approach consists of two separate classifiers and one ensemble of the two(Cls. 1, Cls. 2 and Ensemble respectively) that were used for the classification tasks. The parameters of the models such as the mean and the covariance matrices of the Gaussian distributions, the covariance matrix of the Gaussian kernel used in Cls. 2 or ρ in Equation 1, were all directly estimated from the training data. The regularization parameter λ of the Gaussian kernel, $K_{Gauss}(.,.)$ was preset to 10^{-9}. We also measured the correlation between the predictions from Cls. 1 & Cls. 2 to see how often they predicted the same label.

We compared our approach with several baselines for both the classification tasks. Firstly, we explored the efficacy of unimodal classifiers, using features from each of the 5 feature groups separately and the popular polynomial(poly) kernel Support Vector Machine(SVM) [9] as the classifier(Audio, Text/Doc2Vec, Facet, Gavam, Okao). This also allows us to test the effectiveness of the feature encoding scheme for the text modality, which is our second research objective (Q2).

Next, we explored traditional multimodal classification approaches as baselines, stacking up(early fusion) of features of the modalities followed by a poly kernel SVM classifier(MM Full) and a Naive Bayes classifier(MM Full_NB). The MM Full_NB setting allows us to test if a full conditional independence assumption across all the features is effective or not. We also investigated the effect of dimension reduction techniques. For this, we first performed early fusion of features, then projected the data down to a 5-dimensional space(because Cls. 2 also projects into 5-dim.) using Principal Component Analysis(PCA) and then used a poly kernel SVM classifier for classification(MM Proj).

We also tried several baseline approaches using the features projected in the MVCCA space. First, we used a poly kernel SVM classifier on the data projected on to the MVCCA space, as a baseline(CCA). Further in the last step of Cls. 2, we compute a kernel-similarity ratio for the classification. An alternative classification technique would be to assign the test sample, the label of the majority class of the cluster to which the test sample is assigned. We therefore, also explore this as a baseline(Cls_F). Our final baseline classifier, is the majority vote classifier on the ground-truth(Baseline/ Majority Baseline).

The hyper-parameters for the poly kernel SVMs are determined using 4-fold cross-validation(with 1-partition for validation/ development) on the training data for each of our experiments.

Moreover, we also performed two sample t-tests between our proposed approaches and the baselines to test for statistical significance in classification performance.

For task Q3, where we intend to explore what non-verbal characteristics contribute most significantly to the performance of our model, we ranked the non-verbal features, in their original feature space, by their Information Gain(IG) scores [9]. A more discriminatory feature is expected to have a higher IG score. We report the most significant feature from each feature group for both "passion" and "credibility".

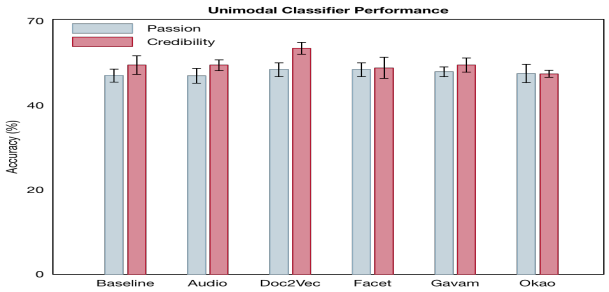

Figure 4: **Accuracies for different Unimodal Classifiers.**

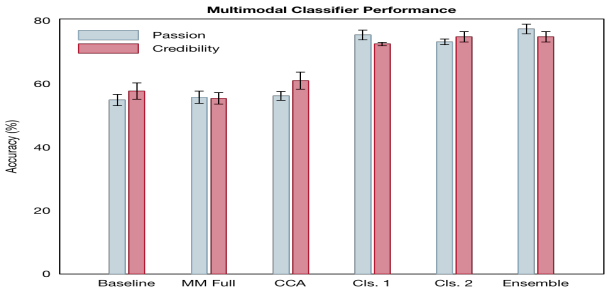

Figure 5: **Accuracies for different Multimodal Classifiers.**

6. RESULTS AND DISCUSSION

6.1 Classification Experiments

6.1.1 *Unimodal Classifiers*

The performance of the unimodal classifiers (see Table 1) show that the classifier for the text modality performs either better or no-worse than the other unimodal classifiers for both credibility and passion. Since all the unimodal classifiers use the same poly kernel SVM classifier and differ only in the features used, we therefore conclude, that our semantic-similarity based feature encoding for text is the most effective amongst all cues individually, for our recognition tasks. This, presents a probable solution to Q2. Figure 4 presents a visualization of the performance of the unimodal classifiers.

6.1.2 *Multimodal Classifiers*

The major observation that emerges from the classification results of the multimodal classifiers for both passion and credibility, is that combining cues from multiple modalities, in a judicious fashion, adds more discriminatory power as compared to considering cues from just one modality(Table 1). We observe this from the performance of our proposed multimodal approaches(Cls. 1, Cls. 2 and Ensemble), which significantly outperform all unimodal baseline classifiers. Table 1 also shows the significance levels for the classification results. Further, the Pearson's correlation measure between the predicted labels of Cls. 1 and Cls. 2 across the 5-folds,

was 0.2028 for "passion" and 0.319 for "credibility". Thus the two predictions were only moderately correlated, thereby suggesting that Cls. 1 and Cls. 2 exploit somewhat different patterns in the data for inference.

Our experiments reveal that not all techniques of combining cues from multiple modalities are equally effective.

In the MM Full setting we perform an early fusion of all the features, in an attempt to model the dependencies between the features of the various feature groups and then classify using a SVM. Again MM Full_NB, presents the complete contrastive setting where we classify using the Naive Bayes classifier, thereby assuming full conditional independence. Our experiments show that neither approaches are ideal for classification, in our context. A sweet-spot between these two extreme assumptions, where we assume conditional-independence between just the feature groups in the combined feature space is more effective. The performance of Cls. 1 compared to MM Full or MM Full_NB shows this. Furthermore, to assuage concerns regarding the issue of high dimensionality of the early fusion feature space leading to over-fitting, we used a PCA baseline(MM Proj). However, it also provides no significant performance improvement because PCA treats all the features from the same perspective. It is incognizant of the fact that the concatenated set of features is a heterogeneous mixture of signals from multiple sources, each with different levels of noise in them. It gives equal importance to all the features and thus, fails to combine the cues effectively.

Another different approach to modeling multimodal data is to explicitly capture the dependencies between every feature group but in a space of few dimensions. MVCCA attempts to address this concern. For our experiments, this results in atleast a marginal gain in performance, over multimodal classification techniques that rely on a modality-unaware dimensionality reduction. We see this when we compare CCA, Cls_F and Cls.2 approaches to MM Proj (atleast a gain of 0.67%).

We also investigated the role of clustering and non-parametric classification techniques in this correlated space. Our experiments divulged that such an approach of clustering followed by classification by assigning the majority class of a cluster(Cls_F) is almost at a-par with a kernel SVM classifier that uses the features directly from the correlated space(CCA). Here we must note that a kernel SVM, is able to capture non-linear patterns in the feature space through the use of a kernel, which is an inherent weakness of the Cls_F model where we simply assign the majority class. Once this weakness is mitigated in our clustering-based model(Cls. 2), we notice a statistically significant performance boost over other unimodal/multimodal baseline classification techniques. This is observable when comparing Cls. 2 with other approaches.

Finally, the linear ensemble of Cls. 1 and Cls. 2 allows us to combine their respective inferences on the data. This brings together the best of both worlds, giving us a performance that is either better or no worse than Cls. 1 or Cls. 2 individually and is better than all other baselines. The parameter ρ allows us to prefer one classifier over the other. For our classification experiments with "passion" the mean ρ across the 5-folds was 0.7280, suggesting the ensemble favored Cls. 1 over Cls. 2. Indeed, even in terms of the classification accuracy on the test-set, for "passion", Cls. 1 is clearly better. For "credibility", the mean ρ across the 5-folds was 0.2040, suggesting the ensemble's tilt towards Cls. 2. Again this in agreement with the results on the test set. We are thus, able to answer Q1.

Figure 5 presents a visualization of the performance of the proposed multimodal classification approaches and a selected set of multimodal baselines.

6.2 Feature Analysis

Our experiments on analyzing the discriminatory power of the non-verbal behaviors of the speaker, our third research goal, brings several interesting observations to light. Tables 2,3 list the most discriminative feature from each category based on their IG scores for "passion" and "credibility" respectively. The higher the score the more discriminative is the feature.

Table 2: **Most discriminative non-verbal descriptors of Passion as measured by Information Gain**

Modality	Statistical Functional	Feature	IG Score
Audio	Standard Deviation	MFCC_5	0.0400
Facet	Standard Deviation	AU 28	0.0200
Gavam	Standard Deviation	Ang. D_Z	0.0100
Okao	75^{th} Percentile	Gaze_LR	0.0311

From the acoustic modality we find the lower(5^{th} or the 6^{th} order) MFCC coefficients to be significant. This is in line with our expectations, since they have denser resolutions and are more robust to noise [22].

From amongst the descriptors obtained from Facet, Action Unit(AU) 28 assumes significance.

Table 3: **Most discriminative non-verbal descriptors of Credibility as measured by Information Gain**

Modality	Statistical Functional	Feature	IG Score
Audio	Standard Deviation	MFCC_6	0.0025
Facet	Standard Deviation	AU 28	0.0046
Gavam	Standard Deviation	Ang. D_Z	0.0032
Okao	Standard Deviation	Gaze_UD	0.0009

The most significant descriptor from Gavam is Angular Diplacement around Z-axis. This corresponds to in-plane head motion and is suggestive of the more expressive gesturing of "passionate", engaging and "credible" speakers.

Finally, we surmise that the most differentiating descriptor from OKAO, namely Gaze deviations up/down(UD) or left/right(LR) hint at the wavering engagement(eye contact) with the audience that is characteristic of non-passionate or inexperienced speakers.

7. CONCLUSIONS & FUTURE WORK

In this work, we attempted to answer 3 research questions in the context of classifying reviewer "passion"(*Pathos*) and "credibility"(*Ethos*). As an answer to the first, we proposed two multimodal classifiers that made polar-opposite assumptions, and both were shown to be effective independently and a linear ensemble of the two was shown to hold promise. We thus, conclude that indeed multimodal cues add discriminatory power compared to just unimodal cues. However, the technique of fusion of these cues plays a pivotal role for achieving good performance. An analysis of the unimodal classifiers showed that the most effective unimodal classifier

Table 1: **Classification Performance of our proposed approach and multiple baselines for Credibility and Passion. The numbers in parenthesis "()" indicate the standard error levels. "[*symb*,*symb*,*symb*]" indicates statistical significance levels between Cls. 1, Cls. 2 and the Ensemble respectively and the current baseline classifier. *symb* could be either $\diamond, \triangle, *$ for p-value < 0.001, 0.01 and 0.05 respectively.**

		Passion	Credibility
Majority Baseline		54.55%(1.77) [◇,◇,◇]	57.38%(2.55) [◇,◇,◇]
Unimodal	**Audio**	54.46%(2.00) [◇,◇,◇]	57.39%(1.45) [◇,◇,◇]
	Doc2vec	**56.15%**(1.88) [◇,◇,◇]	**61.98%**(1.62) [◇,◇,◇]
	Facet	56.15%(1.88) [◇,◇,◇]	56.65%(2.89) [◇,◇,◇]
	Gavam	55.54%(1.32) [◇,◇,◇]	57.39%(1.95) [◇,◇,◇]
	Okao	55.10%(2.51) [◇,◇,◇]	55.02%(0.97) [◇,◇,◇]
Multimodal	**MM Full**	55.40%(1.95) [◇,◇,◇]	55.05%(1.79) [◇,◇,◇]
	MM Full_NB	57.04%(2.66) [◇,◇,◇]	57.36%(3.18) [△,△,△]
	MM Proj	55.17%(1.45) [◇,◇,◇]	59.69%(4.59) [∗,∗,∗]
	CCA	55.84%(1.36) [◇,◇,◇]	60.63%(2.70) [△,△,△]
	Cls_F	57.21%(1.64) [◇,◇,◇]	59.70%(0.62) [◇,◇,◇]
	Cls. 1	74.93%(1.51)	72.11%(0.50)
	Cls. 2	72.78%(0.92)	**74.38%**(1.64)
	Ensemble	**76.85%**(1.50)	**74.38%**(1.64)

was the one corresponding to the text modality, where we represented the text in a space that captures semantic similarity and preserves word-ordering. This seemed to be a promising response to our second research question. As a response to the third research question, we looked into what non-verbal cues were discriminatory. Finally we note, that all our algorithms generalized well for both "passion" and "credibility", giving comparable performance.

For future work, we intend to extend our proposed approach for multi-label classification, try it on other multimodal recognition tasks, and explore feature-selection techniques. On the lines of Chatterjee et al. we also intend to investigate the effect of considering meta-information in the problem context, such as reviewer rating of the movie, for the task of classification [10, 11].

Acknowledgements: This material is based upon work supported by U.S. Army Research Laboratory under contract number W911NF-14-D-0005. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Government, and no official endorsement should be inferred.

8. REFERENCES

[1] Facet. url: http://www.emotient.com/cert.
[2] Gensim. url:https://radimrehurek.com/gensim/.
[3] Okao vision. url: http://www.omron.com/technology/core.html.
[4] M. Abouelenien et al. Deception detection using a multimodal approach. In *16th ACM ICMI*, 2014.
[5] E. Alpaydin. *Introduction to machine learning*. MIT press, 2014.
[6] O. Aran and D. Gatica-Perez. One of a kind: Inferring personality impressions in meetings. In *ACM ICMI*, 2013.
[7] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3, 2003.
[8] C. M. Bishop. Mixture models and the em algorithm. *Microsoft Research, Cambridge*, 2006.
[9] C. M. Bishop et al. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.
[10] M. Chatterjee, S. Park, H. S. Shim, K. Sagae, and L.-P. Morency. Verbal behaviors and persuasiveness in online multimedia content. *SocialNLP 2014*.
[11] M. Chatterjee, G. Stratou, S. Scherer, and L.-P. Morency. Context-based signal descriptors of heart-rate variability for anxiety assessment. In *39th IEEE ICASSP*, 2014.
[12] H. Hotelling. Canonical correlation analysis (cca). *Journal of Educational Psychology*, 1935.
[13] G. A. Kennedy. *On rhetoric: A theory of civic discourse*. OUP, 1991.
[14] J. R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 1971.
[15] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *NIPS*, 2003.
[16] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, 2014.
[17] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *17th ACM SIGIR*, 1994.
[18] T. M. Mitchell. Machine learning. wcb, 1997.
[19] G. Mohammadi et al. Who is persuasive? the role of perceived personality and communication modality in social multimedia. In *15th ACM ICMI*, 2013.
[20] L. Morency et al. Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In *8th IEEE FG*, 2008.
[21] N. Morgan. How to become an authentic speaker, harvard business review, url: https://hbr.org/2008/11/how-to-become-an-authentic-speaker.
[22] S. Park, H. S. Shim, M. Chatterjee, K. Sagae, and L.-P. Morency. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th ACM ICMI*, 2014.
[23] Y. Song, L.-P. Morency, and R. Davis. Multimodal human behavior analysis: learning correlation and interaction across modalities. In *14th ACM ICMI*, 2012.
[24] R. Subramanian, Y. Yan, J. Staiano, O. Lanz, and N. Sebe. On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions. In *Proceedings of the 15th ACM ICMI*, 2013.
[25] J. Vıa, I. Santamarıa, and J. Pérez. Canonical correlation analysis (cca) algorithms for multiple data sets: Application to blind simo equalization. In *13th EUSIPCO*, 2005.
[26] Y. T. Zhuang et al. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *AAAI*, 2013.