Cognitive Science II

Group 8

Multi-Modality Research in IT & Cognition

Automatic Persuasion Recognition

Vejleder: Patrizia Paggio & Costanza Navarretta

Institut for Nordiske Studier & Sprogvidenskab. June 16, 2023

Group Members

Name	KUid
Jonas Kjeldmand Jensen	xgv866
Helena Björnesjö	zvw935
Tudor Laurentiu Dascalu	hpj557

Abstract

In this paper, we present a multi-modal approach to investigating, analysing and interpreting persuasive communication through the lens of machine learning.

We take point of departure from the inspection of critical features of the Persuasive Opinion Multimedia (POM) dataset, we hypothesize that; (1) speaker persuasion can be predicted using text and audio features, by deploying supervised machine learning methods and (2), that bi-modal classification provide better prediction accuracy, as opposed to uni-modal classification for two specific modalities; text and audio.

By collating (fusing) the text and audio modalities, we achieve an accuracy score of *71.68%*, utilizing a bi-modal Random Forest Classifier. However, certain provisos must be considered in regards to the research design chosen, and its concurrent effects on generalisability of the results presented.

Division of Labour

For this project, we deem that every group member have contributed equitably. Every member has taken active part in every step and in every part of the project - from programming, to literature research, to report writing etc. The delightful interdisciplinary nature of the people following IT & Cognition entail a plethora of backgrounds. One in which each member has played to his or hers strengths. Albeit not excluding other group members from actively engaging in tasks where they are more novice. Thus creating an inspiring culture and well-suited environment for mutual learning among peers.

Therefore, we deem it most fair and solidary if the actual paper is graded as a piece of collaborative effort, rather than each individual co-author.

Program Implementation

For this report, we have composed a Python code script. The script is publicly available to anyone via **github**. The script contains everything necessary for interested readers to go and explore, try and test our results for themselves themselves.

Also, the raw files for the entire dataset are made public to anyone by its original authors [22]. The raw files are available through this **link**.

1 Introduction

Cutting through the continuously increasing amount of online content and making an impact is important to many stakeholders. Companies want their products to stand out, politicians and influencers want their voices to be recognised by the many [4, 36, 30]. The purpose of persuasive communication can be seen as an intentional attempt at shaping, supporting or transforming a receiver's current opinion in favour of a sender's position [32, 7, 35, 39, 37].

Social media platforms present a plethora of video and audio content. Commercials, product reviews, political speech, and empowerment speech are examples of types of content for which we suggest that persuasion, considered as a social cue informing recipient response, could be a potential predictor of efficiency of influence [36]. Developing computing frameworks for automatic persuasion recognition is therefore of relevance [28].

In the related fields of personality and emotion recognition, features surface in multiple modalities, such as prosody, intensity and speech activity in the audio modality, and keywords and structural grammatical patterns in the text modality, as well as various visual properties [28, 33, 21]. Much of the existing research on understanding the nature of online behaviour, emotion and persuasion have not worked with multiple data sources of different modalities [14, 19]. However, research on multimodal sentiment analysis, and emotion & personality recognition has demonstrated that prediction performance improves significantly with the integration of different modalities, as compared to measuring the effects of individual modalities separately, suggesting that the whole is indeed greater than the (sum of its) parts in this field [8].

For this project, we present a machine learning approach to understanding, analysing and predicting the persuasiveness of multimedia content. To this end, we study the Persuasive Opinion Multimedia (POM) dataset. The dataset is comprised of three modality measures: *visual*, *audio* and *text* [6, 22]. For this project however, we will only focus on two modalities, namely audio and text.

2 The Dimensions of Persuasive Communication

Already in the *Rhetoric*, Aristotle argued that effectual rhetoric communication depends on *ethos*; the credibility of the speaker, *pathos*; the attitudinal and emotional disposition of the recipient, and, finally, *logos*; the content as well as the form, or structure of the argument set forth, ultimately intended at demonstrating something to be the case [26, 5].

Sidestepping the intricate details of the three concepts as presented in the *Rhetoric*, we may view them as tools for persuasion, the act of influencing a recipient so as to adopt some attitude or belief, by means of employing certain communication strategies [10].

Humans are not purely rational beings. We don't evaluate information only in the form of Aristotelian logical syllogism; the immediate deduction of a proposition from a set of premises. More likely, we blend rational, contextual and emotional information when communicating a standpoint, when forming attitudes or when taking actions. Findings from the emotion research is thus likely to be informative of persuasion [33, 28, 22, 25].

It is believed that affective arousal impacts perception of communication signals [39], and empirical work has shown that affective information impacts recipient attitude formation as a result of being presented with emotionally charged persuasive/non-persuasive speech. In this sense, affective expressions convey information in communication settings [32].

For audio, prosody (including intonation, pitch and rhythm) is widely held to correlate with emotional states such as distress and anxiety and studies have demonstrated the importance of prosodic features for emotion recognition in speech [28, 17, 29]. With regards to text, emotion words, word polarity, punctuation, and letter capitalisation are examples of properties that are informative of emotion [28].

Park et. al. [22] who first introduced the POM dataset, suggested that persuasion and personality characteristics may also be related. They found that their predictions for speaker persuasiveness showed significant positive correlation between persuasiveness and four of the Big Five personality traits (extraversion, agreeableness, openness, conscientiousness), whereas neuroticism showed significant negative correlation with persuasion [8].

Also the field of Human-Computer Interaction has since long taken great interest in discovering viable approaches to integrate perspectives from personality psychology [18].

The study of persuasion then, is interesting in that it is informed by and brings together multiple research domains.

3 Towards a Research Agenda

While personality psychology has spurred much excitement in the field of Human-Computer Interaction, multiple issues and challenges remain unsolved in the endeavour to connect these fields of research [23, 34, 18]. The affective computing field has largely focused on audio-visual feature fusion, leaving the contribution of text underexplored [25, 15, 6]. Interestingly, Poria et al. [25] found that text and audio features displayed better classification performance with respect to unseen emotions and sentiments than visual features did.

In personality computing, text has been an important descriptor to exemplify how users externalise their personality through the uses and affordances of technology [34]. Text is as of yet still a major medium for communicating opinions online and this is reason to address the text modality in persuasion computing.

As interesting as it would be to investigate links between persuasion and qualitative psychological and cognitive properties in order to identify speaker characteristics associated with persuasion, or emotional recipient responses to persuasive/non-persuasive speakers, such investigations fall outside the scope of our endeavour. Our method focuses on persuasion recognition based on dataset annotations of perceived persuasion intensity.

For the purposes of this project, we ask if speaker persuasion can be predicted from text and audio descriptors using machine learning classification models. In response to this question, we hypothesise that (1) speaker persuasion can be predicted using text and audio features, by deploying supervised machine learning classification models, and (2) that bi-modal classification yields higher prediction accuracy than uni-modal classification for the two respective modalities text and audio.

4 The POM Dataset

The Persuasive Opinion Multimedia dataset was first introduced by Park et. al. [22]. The corpus consists of 1.000 movie review videos obtained from the website expotv.com. It has annotations for personality, sentiment and high-level attributes. The original idea behind constructing the POM dataset was to enable exploration of persuasiveness. As the dataset is publicly available, a multitude of research teams have experimented with the POM dataset [20, 38, 27, 9].

The dataset corpus is multi-modal, and encompasses three modalities: *visual*, *audio*, and *text*. Text consists of transcribed files of the the visual/audio modalities.

Each review comes with a sentiment star rating of a movie made by the reviewer, ranging from 1 (least positive) to 5 (most positive), reflecting speaker opinion polarity. It is suggested by Park et al. that a 5 star review is likely to correspond to an attempt to persuade the recipient in favor of the movie, and the other way around for 1 star reviews [22]. Out of the total 1000 videos, 500 had a 5 star rating. There are 216 videos with a 1 star rating. To obtain same size persuasion data subsets, 284 2 star ratings were added to the 1 star ratings.

Annotation was done through crowdsourcing on Amazon Mechanical Turk. In total, 50 native English speakers were recruited as annotators.

The high-level attributes labels consist of a set of properties that Park et al. suggest to be associated with persuasion: credible, entertaining, expert, passionate, confident, humorous, professional-looking, physically attractive, voice pleasant, and vivid. These attributes were measured on a Likert scale, 1 (least descriptive) to 7 (most descriptive) [22].

For persuasion, annotators assigned an integer ranging from 1 (very unpersuasive) to 7 (very persuasive). Three rounds of annotations were carried out and the resulting labels consist of the average of the numbers from each round. For our project, we used the persuasion scores to classify each video as being persuasive or non-persuasive. Rather than splitting the dataset using the mean score (3.5), we consider a video to be persuasive if its average score is higher than 5. On the other end, a non-persuasive video has a score that is lower than 3. The majority of the movie reviews were labeled between 3 and 5 on the persuasion scale. Hence, the dataset used for training and testing our models comprises 384 videos.

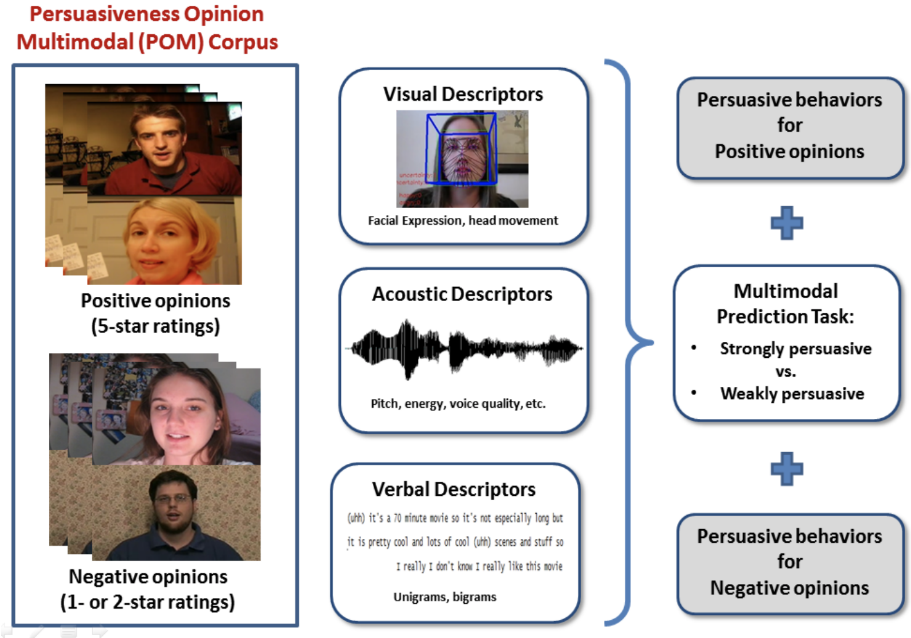
Persuasiveness Opinion Multimodal (POM) Corpus
Positive opinions (5-star ratings)
Negative opinions (1- or 2-star ratings)
Visual Descriptors
Facial Expression, head movement
Acoustic Descriptors
Pitch, energy, voice quality, etc.
Verbal Descriptors
(uhh) it's a 70 minute movie so it's not especially long but it is pretty cool and lots of cool (uhh) scenes and stuff so I really I don't know I really like this movie
Unigrams, bigrams
Persuasive behaviors for Positive opinions
Multimodal Prediction Task:
Strongly persuasive vs. Weakly persuasive
Persuasive behaviors for Negative opinions

Figure 1: Figure showing POM corpus. Taken from Park et. al. [22].

5 Methods

In this section we present the methods utilised in this project, prevailing methods that went into the end-design. First, we present some preliminary assumptions about our data and how we use it. We then proceed to present the steps of our method framework, feature extraction, feature selection, and classification.

5.1 Preliminaries

For the purpose of this project, we restrict the scope of our investigations to the text and audio modalities. First, we extract features from the two modality channels. Next, we fuse the separate feature vectors. For this step, we utilise a supervised Random Forest model and a Gradient Boosting model for classification. We test our bi-modal classifiers on the POM dataset. Moreover, we compare the classification accuracy of the bi-modal classifiers with uni-modal audio and text classifiers.

In our framework, recipient disposition is accounted for insofar as the data is annotated. Labels attributed to a video are taken to reflect viewer disposition for action, the act of rating the speakers' level of persuasion on a scale from 1 to 7. As a working assumption, we suggest this results from temporarily held attitudes corresponding to impressions of the movie reviewer formed as a result of watching a particular video.

We use the POM persuasion labels. Persuasion intensity is measured indirectly in terms of annotator labeling. For our purposes, the averaged label obtained from three annotation rounds assigned to a video is treated as a benchmark for the true level of persuasion that the presenter of that video displays [22].

5.2 Feature Extraction

For audio feature extraction, we used the CMU-Multimodal-SDK library, which provided us with audio features processed using the COVAREP voice analysis repository. Each file is represented as a time series with 43 dimensions, that include information such as pitch, 12 Melfrequency cepstral coefficients and voiced/unvoiced segmenting features for every 10ms of audio input [31].

For text feature extraction, we have implemented two methods. From these, the statistical term-frequency-inverse document frequency (TF-IDF) measure for term relevance, gave the highest accuracy on persuasive/non-persuasive text prediction. TF-IDF assigns words that appear in multiple corpora documents low relevance. Contrarily, words that appear multiple times in one document but are not present in several other documents are assigned high relevance. We also implemented a Doc2Vec embedding using a 128-dimensional Google news model for transfer learning. Doc2Vec is a dense neural network that learns word embeddings based not only on the immediate surrounding words for some word, but also on the document in its entirety. A document is represented as the average of word embeddings for randomly sampled document words.

5.3 Feature Selection

Text feature selection was based on the TF-IDFvectorizer, which tokenises the input text documents and yields a feature matrix representation. We tried different maximum top number of features to be selected from a document (100, 500, 1000). For our final implementation, this was set to 500. After trying different values (1, 5, 10), we set the minimum threshold for term occurrence count in a document to 5. Similarly, we ignored terms that exceeded a maximum threshold (50%, 70%, 90%) for number of documents that the term appeared in. In our final implementation, this was set to 70%. In addition, we experimented with both unigrams and bigrams and ended up using unigrams. Using the NLTK library, we also removed high-frequency stop-words (such as 'an', 'the', 'and', etc.).

For audio feature selection, we reduced the dimensionality of the COVAREP time series audio representation and made the data uniform across all audio files. This was done by computing the mean, standard deviation, minimum, maximum, range (maximum - minimum), and skewness of the time series data across all features [20].

5.4 Models for Classification

As a baseline model, we implemented a *DummyClassifier*, which utilises the *most_frequent* method to classify data points. In order to retrieve opinions that were 'truly' persuasive/non-persuasive, we removed all middle label values (y<3 y>5). In this way, we deem that with a higher degree of certainty, we can assess a review to be an instance of persuasive communication or not. Thus, making the problem into a binary classification issue. Also, since the data is balanced (roughly same number of persuasive and non-persuasive content), the classifier scores low on accuracy during cross validation.

We used the highly optimised *RandomForestClassifier* implementation from the sklearn library, which sub-samples and averages the dataset and fits decision tree classifiers to the sub-samples. We experimented with different best split estimators: [100, 300, 500, 1000].

We also used the highly optimised *GradientBoostingClassifier* implementation from sklearn which produces an ensemble of decision trees in an iterative process and minimises the differential loss function [1]. We experimented with different best split estimators: [100, 300, 500, 1000].

For our dense neural network, we used TensorFlow Keras Sequential model. Again, we experimented with different parameter values. The final implementation of the model includes two hidden Dense layers with logistic sigmoid activation functions, one Dropout layer and a output Dense layer, also with a sigmoid activation function. We use binary cross entropy as loss function. The Dropout layer has a dropout rate of 20% to reduce the risk of over-fitting on the training data.

In order to evaluate our models, we used 5-fold Stratified Cross Validation. The reason why we opted for this method is because the size of the data set is relatively small, and we want to ensure that each fold includes an (approximately) equal number of persuasive and non-persuasive movie reviews.

5.5 Feature Fusion for Bi-Modal Classification

For bi-modal classification of the persuasion data, we combined the two modalities, text and audio, using two different approaches: early and late feature fusion.

For early feature fusion, we simply concatenated the audio and text feature vectors before training the Random Forest classifier, Gradient Boosting classifier and Neural Network models.

Our approach to late feature fusion consists of first having trained the models for audio and text feature vectors separately. Next, we computed the class probabilities (e.g persuasive: .30, non-persuasive: .70) for model prediction on the data test set. Finally, we computed the average of the probabilities and selected the class with the highest average probability.

6 Results

In the following section, we present the results from experimenting with uni-modal text and audio classifiers and fused bi-modal classifiers. We also show the results from the *DummyClassifier*, which we used as a baseline for performance comparison.

6.1 Uni-Modal Text Classification

This table shows the accuracy scores for different methods that we tested. As shown, TF-IDF feature extraction with Random Forest classification produced the best results with a mean accuracy score of 69.02%.

Method	Features	Mean Accuracy
Dummy Classifier	Most Frequent	51.30%
Random Forest	TF-IDF	**69.02%**
Gradient Boosting	TF-IDF	63.54%
Neural Network	TF-IDF	62.25%
Neural Network	Doc2Vec	62.77%

Table 1: Performance of classifiers using text features

6.2 Uni-Modal Audio Classification

This table shows the accuracy scores for the three classification models that we tested on the COVAREP audio feature vectors. Again, the Random Forest implementation produced the best result with a mean accuracy of 65.63%.

Method	Features	Mean Accuracy
Dummy Classifier	Most Frequent	51.30%
Random Forest	COVAREP	**65.63%**
Gradient Boosting	COVAREP	60.94%
Neural Network	COVAREP	52.35%

Table 2: Performance classifiers using audio features

6.3 Bi-Modal Classification

In this table, we present the results obtained from fusing the text and audio modalities. We see the results from both the early fusion and late fusion methods that we implemented for bi-modal classification.

We only present early fusion for the Neural Network method, since its performance was close to the *DummyClassifier* baseline. The late fusion Random Forest model rendered the best results with a mean accuracy score of 71.61%, well above baseline accuracy.

Method	Features	Fusion	Mean Accuracy
Dummy Classifier	Most Frequent	-	51.30%
Random Forest	COVAREP, TF-IDF	Early	66.67%
Gradient Boosting	COVAREP, TF-IDF	Early	67.44%
Neural Network	COVAREP, TF-IDF	Early	52.62%
Random Forest	COVAREP, TF-IDF	Late	**71.61%**
Gradient Boosting	COVAREP, TF-IDF	Late	62.50%

Table 3: Performance of classifiers using audio and text features

7 Discussion

The discussion section takes its point of departure with a tour of our obtained results, explaining to what extend and with which limitations we were able to address our two hypotheses. Next, we inaugurate critical reflections on our method selection, while contrasting these against alternative methods we either directly or indirectly experimented with. Further, we debate to what level our results are generalisable and the plethora of challenges that arise from utilising novel data. Lastly, we outline potential point of departures for future work, i.e. how we could have proceeded to further develop on our programming strategy, or how potential alternative outcomes might have influenced this research project differently.

7.1 Connecting Our Results Back to the Hypotheses

The Random Forest model using COVAREP and TF-IDF feature vectors predicts persuasive content with an upper accuracy of 75% and mean accuracy over 5 folds of 71% (Table 3), which is considerably higher than the baseline accuracy. This supports our hypothesis (1) that speaker persuasion can be predicted using text and audio features, by drawing on machine learning classification methods. With respect to our second hypothesis, we also see that the bi-modal classifier performed better than both of the uni-modal classifiers. The uni-modal audio classifier performed worse than the uni-modal text classifier. Moreover, the models have low computational cost and relatively high transparency, which enables broad application.

The late-fusion bi-modal approach outperformed both uni-modal models. It is interesting that early fusion of text and audio features leads to lower accuracy on the validation set, as opposed to uni-modal text classification alone. We believe the reason for this is that the dimensionality of the two feature vectors is not proportionate (text - 500, audio - 258), resulting in some essential text descriptors loosing weight. Moreover, TF-IDF matrices are sparse, which means that the audio features contribute more in the final say.

7.2 Feature exploration

To visualise some of the most informative text features for persuasiveness, we used Pearson's chi-squared test and plotted the first fifteen in figure 2 below. [2]

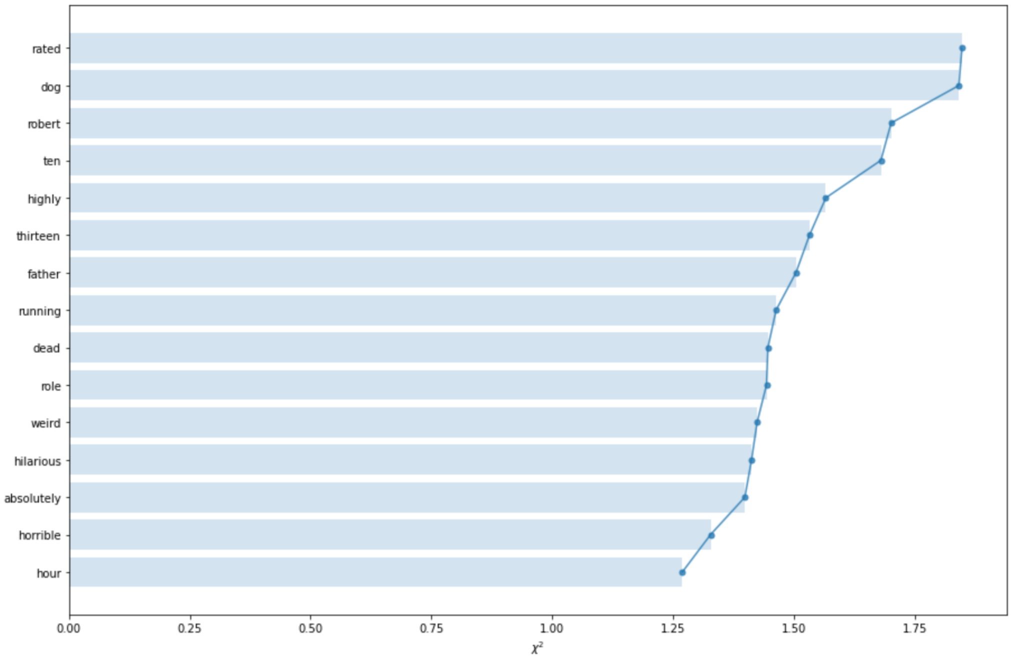

Figure 2: Figure shows the χ^2 scores for the 10 highest rated words in ascending order.

In figure 2, we see words such as 'father', 'dead' and 'thirteen' which, in the context of movie reviews, might be informative of movie content. We also see adjectives such as 'hilarious', 'horrible',

’weird’ and ’highly’. Similarly, among the ten best words, we have ’great’ and ’see’. These adjectives are likely to reflect reviewer sentiments. Several of them display more intensity than more neutral adjectives such as ’okay’ or ’acceptable’. This is in line with what could be expected of persuasive speech, since a persuasive speaker is likely to emphasise their opinions of the topic at hand.

7.3 Critical Reflections on Our Methods

During the data exploration phase, we noticed that most of the videos were labeled as ranging in persuasiveness from 3 to 5. In order to test our hypotheses, we decided to work with highly persuasive and non-persuasive content and removed movie reviews with labels greater than 3 and smaller than 5. This is sensible as videos labeled as neutral would be expected to contain less persuasion specific descriptors.

For audio feature extraction we used all 43 dimensions constructed with COVAREP. The uni-modal audio classifier performed worse than both the uni-modal text classifier and the bi-modal classifier. Quite possibly, this could be due to the fact that we did not approach the issue from a time-series perspective. Perhaps a persuasive speaker opens and ends a review on a particularly persuasive note. A way to address this suggestion would be slicing the data to extract features from shorter time intervals. For example, segments at the beginning or end of a video might be richer in persuasive content. We speculate, this plan might have resulted in more persuasion-specific features for the audio modality.

Ever evolving social media abbreviations are often not captured by word embedders. Video text transcripts also cover speech fillers such as ’uhuh’ or ’mmm’. Such phenomena are often overlooked by word embedders. It is likely that the Google news dataset that is used for transfer learning in our Doc2Vec implementation does not cover such expressions. This might have affected the performance of the model and thus the accuracy score we see for the Doc2Vec implementation. This could be addressed by manually identifying fillers for exclusion from the POM text transcripts. Those fillers could then be excluded from the text files using detailed hand-written *regular expressions* as part of preprocessing the text data. In general, the quality of a neural network depends on preprocessing, as well as on fine-tuning of the network parameters; such as choice of activation and loss functions, batch size, number of epochs, and number of hidden layers. Although we experimented with different parameter values, it is possible that the tuning of these parameters could be further improved in our implementation.

We saw in the results section that the uni-modal text classification yielded better results than audio classification did. This could be addressed by assigning more weight to the classifier trained on text data. Another option could have been to use a different fusion strategy, such as a hierarchical approach, which have been successful in other settings [6]. In addition, the use of automatic methods for hyperparameter tuning, such as RandomizedSearchCV from sklearn [3], could improve the decisions of the best performing models (Random Forests).

7.3.1 Annotator Bias

A thought listing technique was utilised during data gathering to control for annotator sentiment towards - measured as interest in - the movie under review, on a scale from -3 (little interest) to +3 (much interest) [22]. It is possible that strong interest in a movie affects the perception of a reviewer, and similarly that a lack of interest in, or a direct opposition to, a movie might result in a lower annotation score. In a sense, this point could be extended to the question of: ’Was the annotator persuaded?’. It is possible that what was annotated was not primarily the movie reviewers’ level of persuasion, but the movie itself. These sentiment labels could have been used in our approach as well, to control for possible influence from attitude towards a movie interfering with persuasion prediction.

7.3.2 Limitations to TF-IDF

When evaluating our winning method, obvious advantages present themselves. The TF-IDF method is easy to compute. It provides a basic metric, which enabled us to extract the most descriptive words in each document. To knot the two previous benefits, the TF-IDF method is then also computationally inexpensive when searching for similarities between two documents. In this way, this relatively simple algorithm is able to match words in the query to the document relevant to query.

However, this method does carry certain disadvantages. Representing text in numeric form has shown to be a challenging feat in machine learning [12]. Different methods exists, but their performance are mediocre at best.

For text vectorisation, TF-IDF is based on a Bag-of-Words (BoW) philosophy. The BoW approach however, will lose many of the nuances you would like to have in a semantically rich word representation. This implies that TF-IDF is unable to capture a word's position in the text, the semantics of the text and co-occurrence of words across all the documents in the corpus. To add to this, the model assumes that the counts of different words provide independent evidence of similarity, and it makes no use of semantic similarities between words. When considering all the above arguments, it can be reasoned that the TF-IDF is a model which is limited to only capturing low-dimensional, latent representations.

For these reasons, the TF-IDF model is most applicable in cases where you are purely working on the lexical level and not trying to deduce some higher semantic meaning. This lead us on to Doc2Vec as a more refined method for executing text vectorisation [13, 16]. However, this method would show to produce inferior results compared to TF-IDF. We speculate that one reason for this might be due to the nature of language in social media text, as it is often riddled with shortcuts, abbreviations and misspellings.

7.4 Generalisability - The Challenge from Novel Data

We have tested our classifiers on unseen data. However, all the data used for the testing/training split is from the POM dataset. This raises some questions regarding the generalisability of our models. Due to time constraints, we did not test our models on other data.

Using data from other sources could provide a stress test for the viability of our model. We considered extracting YouTube video content for further testing, using some metric such as shares or likes as a benchmark for comparison. If such a test would have resulted in a significant positive correlation between predicted persuasiveness and high outreach metric score, and between non-persuasive prediction and low outreach metric score, this would have provided some support for the generalisability of our model. Time constraints aside, we also took into account the difficulty of controlling for other factors influencing YouTube metrics, such as fame and popularity of speaker/channel. Depending on the data, such properties might have led the model to predict persuasion in videos where fame is the main explanatory factor. In such cases, we would be unable to attribute results to persuasion in particular.

In general for (supervised) automatic modeling, the labeled data is different from the target data. This does not primarily pose a problem for the validity of the internal workings of a model, but it can place constraints on the applications of it. Political discourse is different from that of holiday commercials, and left-wing political discourse may be different from right-wing political discourse [7, 4]. This entails that features that are indicative of persuasion in one domain, might not predict persuasion accurately in another domain. Video material in English directed at an L1 American audience might not generalise to L2 audiences in Denmark or Japan. The impact of persuasive communication is then likely to vary with the social and cultural context that it exists in [11].

7.5 Future Work

We suggest future work extends the framework presented in this paper to include visual descriptors, which carry significant information about bodily gestures and facial expressions. We cogitate that speakers intending to convince their audience of a message are likely to use facial expressions and gestures to increase their chance of persuading their receiver.

Regarding text feature extraction - an interesting approach would be to combine TF-IDF with the Linguistic Inquiry Word Count (LIWC) dictionary, which cover psychological, social, and affective word categories. Grouping words into such categories might provide features that are more informative with respect to persuasive communication.

We considered experimenting with a transformer language model on the text data, such as the state-of-the-art Bidirectional Encoder Representations from Transformers (BERT). BERT, and other bi-directional embedders, are fit for contextual sentence modeling. BERT is powerful, but computationally expensive and highly complex. A somewhat similar approach using glove word embeddings could be used. Our Doc2Vec implementation also accounts for context, but operates with a simpler structure than some of the more recent embedders. BERT, glove or other available pre-trained embedders might improve classification performance.

For feature fusion, alternative methods could prove fruitful to test, such as hierarchical approaches as previously touched upon. One strategy that is present in the multi-modal sentiment analysis literature is to use Long Short-Term Memory (LSTM) for modeling contextual/temporal dependencies [24].

Returning to previous sections, we encourage investigations of persuasion drawing on emotion & personality research. Further exploration of underlying cognitive mechanisms of the perception of persuasive speech, as well as attributes of persuasive speakers could reveal cognitive properties associated with persuasion. The POM dataset comes with labels for high-level features (such as ”confident” or ”vivid”), enabling comparison between persuasion and some such qualitative properties. Triangulation studies looking at correlations between those properties and the persuasion labels, or comparing the persuasion labels with labels for affective or personality properties, would contribute insights into the domains of computational persuasion recognition and persuasion perception.

As a final remark, aside from previously mentioned possible areas of interest for further study, we suggest persuasion to be interesting and potentially informative with respect to human-computer interaction and mass communication research, as well as being relevant to applied areas such as political forecasting, designing speeches, marketing, and campaigning.

8 Conclusion

We set out to test two hypotheses, and our results provide support for both hypotheses. In support of the first hypothesis, the bi-modal classifier performed markedly better than the baseline. This implies that the bi-modal Random Forest model with COVAREP and TF-IDF feature vectors for audio and text, respectively, makes acceptable predictions of persuasion.

With respect to our second hypothesis, the bi-modal classifier performed better than both uni-modal classifiers did as was expected. The uni-modal text classifier performed reasonably well on its own, which indicates that text features were most informative for prediction. Seeing that the bi-modal classifier still resulted in highest accuracy score supports the implicit assumption that multi-modal modeling improves prediction performance compared to uni-modal prediction.

9 Bibliography

References

[1] 3.2.4.3.5. sklearn.ensemble.GradientBoostingClassifier — scikit-learn 0.23.1 documentation.

[2] sklearn.feature_selection.chi2 — scikit-learn 0.23.1 documentation.

[3] sklearn.model_selection.RandomizedSearchCV — scikit-learn 0.23.1 documentation.

[4] S. Ahmad. Political behavior in virtual environment: Role of social media intensity, internet connectivity, and political affiliation in online political persuasion among university students. *Journal of Human Behavior in the Social Environment*, 30(4):457–473, 2020. Publisher: Routledge.

[5] Aristoteles. Aristotle on rhetoric: a Theory of Civic Discourse /, 1991. ISBN: 9780195064865 Place: Oxford.

[6] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[7] R. I. Bancoș. VISUAL RHETORIC OF POLITICAL DISCOURSE. AN ANALYSIS OF A POLITICIAN'S LOCAL ELECTION CAMPAIGN FOR THE EUROPEAN PARLIAMENT. *Studia Universitatis Babes-Bolyai - Ephemerides*, 60(1):5–24, 2015. Publisher: Studia Universitatis Babes-Bolyai.

[8] E. Cambria. Recent trends in deep learning based personality detection. *The Artificial Intelligence Review*, 53(4):2313–2339, 2020.

[9] M. Chatterjee, S. Park, L.-P. Morency, and S. Scherer. Combining Two Perspectives on Classifying Multimodal Data for Recognizing Speaker Traits. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 7–14, Seattle, Washington, USA, Nov. 2015. Association for Computing Machinery.

[10] W. D. Crano and R. Prislin. Attitudes and Persuasion. *Annual Review of Psychology*, 57(1):345–374, 2006. _eprint: https://doi.org/10.1146/annurev.psych.57.102904.190034.

[11] S. Hall. Cultural studies: two paradigms. *Media, Culture & Society*, 2(1):57–72, Jan. 1980. Publisher: SAGE Publications Ltd.

[12] D. Kim, D. Seo, S. Cho, and P. Kang. Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. *Information Sciences*, 477:15–29, Mar. 2019.

[13] Q. V. Le and T. Mikolov. Distributed Representations of Sentences and Documents. *arXiv:1405.4053 [cs]*, May 2014. arXiv: 1405.4053.

[14] A. C. Lima and L. N. d. Castro. Multi-label Semi-supervised Classification Applied to Personality Prediction in Tweets. In *2013 BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence*, pages 195–203, Sept. 2013. ISSN: 2377-0597.

[15] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria. Multimodal Sentiment Analysis using Hierarchical Fusion with Context Modeling. *arXiv.org*, 2018.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, Sept. 2013. arXiv: 1301.3781.

[17] C. Montacie and M.-J. Caraty. Pitch and intonation contribution to speakers' traits classification. *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 1:526–529, Jan. 2012.

[18] C. Nass and K. M. Lee. Does Computer-Synthesized Speech Manifest Personality? Experimental Tests of Recognition, Similarity-Attraction, and Consistency-Attraction. *Journal of Experimental Psychology: Applied*, 7(3):171–181, 2001. Publisher: American Psychological Association.

[19] C. Navarretta and L. Oemig. Big Data and Multimodal Communication: A Perspective View. pages 167–184. July 2019.

[20] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrusaitis, and L.-P. Morency. Deep Multimodal Fusion for Persuasiveness Prediction. Nov. 2016.

[21] P. Paggio, L. Galea, and A. Vella. Prosodic And Gestural Marking Of Complement Fronting In Maltese, 2018.

[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 International Conference on Multimodal Interaction*, ICMI '14, pages 50–57, Istanbul, Turkey, Nov. 2014. Association for Computing Machinery.

[23] F. Pianesi. Searching for Personality [Social Sciences]. *IEEE Signal Processing Magazine*, 30(1):146–158, 2013. Publisher: IEEE.

[24] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency. Context-Dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[25] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain. Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines. *IEEE Intelligent Systems*, 33(6):17–25, 2018.

[26] C. Rapp. Aristotle's Rhetoric. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2010 edition, 2010.

[27] A. Ravichander, S. Rijhwani, R. Kulshreshtha, C. Nagpal, T. Baltrusaitis, and L.-P. Morency. Preserving Intermediate Objectives: One Simple Trick to Improve Learning for Hierarchical Models. June 2017.

[28] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1):28, Dec. 2018.

[29] M. Sanchez, A. Lawson, D. Vergyri, and H. Bratt. Multi-System Fusion of Extended Context Prosodic and Cepstral Features for Paralinguistic Speaker Trait Classification. *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 1, Jan. 2012.

[30] J. Schwanholz, T. Graham, and P.-T. Stoll. *Managing Democracy in the Digital Age Internet Regulation, Social Media Use, and Online Civic Engagement.* Springer International Publishing, Cham, 1st ed. 2018. edition, 2018.

[31] J. Tang, M.-Y. Kan, D. Zhao, S. Li, and H. Zan. *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II.* Springer Nature, Sept. 2019. Google-Books-ID: Ne2yDwAAQBAJ.

[32] G. van Kleef, H. van den Berg, and M. Heerdink. The Persuasive Power of Emotions: Effects of Emotional Expressions on Attitude Formation and Change. *Journal of Applied Psychology*, 100:1124–1142, July 2015.

[33] S. Villata, E. Cabrio, I. Jraidi, S. Benlamine, M. Chaouachi, C. Frasson, and F. Gandon. Emotions and personality traits in argumentation: An empirical evaluation 1. *Argument & Computation*, 8(1):61–87, Jan. 2017.

[34] A. Vinciarelli and G. Mohammadi. A Survey of Personality Computing. *IEEE Transactions on Affective Computing*, 5(3):273–291, July 2014.

[35] A. Vize. The art of persuasion: analysing environmental media. *Screen Education*, (63):66–70, 2011.

[36] B. E. Weeks, A. Ardèvol Abreu, and H. Gil De Zúñiga. Online Influence? Social Media Use, Opinion Leadership, and Political Persuasion. *International Journal of Public Opinion Research*, 29(2):214–239, 2017. Publisher: Oxford University Press.

[37] D. Yasmina, M. Hajar, and A. M. Hassan. Using YouTube Comments for Text-based Emotion Recognition. *Procedia Computer Science*, 83:292–299, 2016. Publisher: Elsevier BV.

[38] A. Zadeh, S. Poria, P. Vij, and E. Cambria. Multi-attention Recurrent Network for Human Communication Comprehension. *arXiv.org*, 2018.

[39] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, Jan. 2009. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.