Spring 2020

An Study of Sentiment Analysis Methods For Mandarin Chinese

Thalia Celeste Chai-Zhang
Bard College, tz8035@bard.edu

An Study of Sentiment Analysis Methods For Mandarin Chinese

A Senior Project submitted to
The Division of Science, Mathematics, and Computing
of
Bard College

by
Thalia C. Chai-Zhang

Annandale-on-Hudson, New York
May, 2020

Abstract

Sentiment analysis is the study of automated methods of emotion detection in language and speech. It is an area of much active study, and has seen significant progress in improving accuracy over recent years. Many studies however are conducted on datasets of English and other Indo-European languages. We apply well studied methods for sentiment analysis to a Mandarin Chinese data set and compare the accuracies achieved against eachother and other studies. Our findings indicate that methods for improving sentiment analysis accuracy for English may not be as applicable to Mandarin Chinese.

Contents

Acknowledgments

This study would have been able to be conducted without the funding and support of the Lifetime Learning Institute's Seniors to Seniors grant and the Bard Computer Science department. I would also like to thank my advisor Sven Anderson for his continuous support of this research.

1
Introduction

Emotions are a fundamental part of being human and how we experience the world. But emotions are not merely contained within ourselves, they are something we express in our every movement, in every word we say, every expression we make and even how we unconsciously carry ourselves. Humans are social creatures, and we are constantly communicating our emotions out to the world, and reading the emotions of those around us. So it is that when it comes to designing machines and programs to interact with and help people that we must approach the issue of how to automate emotion recognition, also known as sentiment analysis.

While we communicate in many ways, speech is one of the most immediately emotive of them. It is necessary to understand a written language in order to determine the emotional sentiment it carries, but one does not necessarily need to know the exact textual meaning of what someone is saying to know that they are saying it in an angry or sad or happy way. This study is concerned with using sentiment analysis to determine the sentiment of spoken speech Mandarin Chinese without analysing the meaning of the words spoken. Mandarin Chinese, despite being the most spoken language in the world is not often used to study sentiment analysis compared to English and other Indo-European languages (Wu, 2006; Chen, 2018).

We use a number of well studied methods for sentiment analysis with a limited set of well understood speech audio features extracted from a corpus of Mandarin speech utterances, and compare the accuracy of these methods against each other and results from similar studies. We aim to determine how well studied methods for sentiment analysis work for a language less commonly studied by the field.

2
Background

2.1 Sentiment Analysis

Automated sentiment analysis or emotion recognition from human speech has been a field of interest to researchers in Computer Science, Psychology and Cognitive Science for several decades now (Tao, 2005). While overall sentiment classification accuracy has improved over time as more advanced methods for sentiment analysis have been researched and developed, classification accuracy still suffers significantly, especially in classification problems with more than three sentiment classes (Li, 2016; Majumder, 2018). Even more complex models that incorporate video analysis of the speaker or automatic speech recognition to analyze textual feature of the speech have been unable to substantially improve upon this issue of specific sentiment classification, achieving at best approximately 75% accuracy in four or five sentiment classification problems (Majumder, 2018; Zhong, 2012).

Other research has been conducted on improving methods of feature extraction, and selecting more features that are relevant to sentiment analysis from the speech samples. These, however, run into issues of higher dimensionality inherent to using more features, but still achieve overall comparatively high accuracy approaching 80% in a six sentiment classification problem (Wang, 2015). Further research into better processing these higher dimensional features using dimensional reduction methods and more complex decision models has been even more promising, with some approaching 85% accuracy in five sentiment classes (Sun, 2019).

While these advances are encouraging, sentiment analysis remains an area of intense study and difficult challenges. Many of these advances depend on increasing the size and dimensionality of the feature set, which incurs its own issues with decreased generalized performance and increased computational requirements (Gu, 2015). Furthermore, much of the current research literature has primarily studied a limited selection of languages, in particular English and other Indo-European languages (Wu, 2006). Much less research has been done on languages such as Mandarin Chinese, despite being the most widely spoken language globally (Wu, 2006; Chen, 2018), and there is significantly decrease accuracy for sentiment analysis in these languages outside of the main focus of study (Chen, 2018; Sun, 2019).

3
Methods

3.1 Feature Extraction

Feature extraction is the vital first step in sentiment analysis, in which predictor features are collected from the datasets to create the feature set used by sentiment analysis models for training and testing. Feature extraction in this study was conducted using the Praat phonetics speech analysis open-source software, which allowed for easy scriptable extraction of the speech features described below.

3.2 Features

Intensity Features

Intensity is approximated by the sound pressure expressed in decibels relative to the auditory threshold. It is often considered vital for sentiment analysis due to associated between high intensities and happiness as well a panic, and between low intensities and sadness and neutrality (Liu, 2003). The sum, mean, standard deviation, max and minimum of intensity were collected. Timescale vectors of intensity as well as linear regressions of intensity were generated by sampling 10 spoken peaks along the length of the utterance.

Pitch Features

Pitch is measured as frequency in hertz. Pitch is considered a very fundamental feature for sentiment analysis (Poria, 2016), as high pitch is often associated with happiness while low pitch is more often associated with sadness (Liu, 2003). The mean and standard deviation of pitch were collected. Timescale vectors of pitch as well as linear regressions of pitch were generated by sampling 10 spoken peaks along the length of the utterance.

Harmonacity

Harmonacity is the degree of acoustic periodicity, also called Harmonics-to-Noise Ratio (HNR) expressed in decibels, representing the degree of discrete signal-to-noise in a sound. Harmonacity can be used as a measurement of human voice quality, with low harmonacity implying hoarseness of voice (Poria, 2016). The mean, standard deviation, max and minimum of harmonacity were collected.

Speaking Rate

Speaking rate is the frequency of discrete syllables detected by assessing peaks in intensity preceded by dips and discernible changes in pitch, over time. Articulation rate is a similar metric that accounts for periods of significant silence to determine only the speaking rate for times in which speech is detected. There is some controversy over what sentiments higher and low speaking rates are associated with, in particular anger has been shown to on occassion be associated with both high and low speaking rate (Ayadi, 2011).

In total this resulted in a feature set of thirteen discrete features and two timescales for each utterance, a comparatively small number compared to the dozens or hundreds used by some studies (Chen, 2012; Majumder, 2018). These features were extracted from the 20400 utterances of the Mandarin Affective Speech Corpus (MASC) (Wu, 2006).

3.3 Experiments (Comparison of Sentiment Analysis Models)

A total of five different classifier methods were tested in this study: Linear Discriminant Analysis, Logistic Regression, a Random Forest Classifier, Support Vector Machine and a two layer decision tree Support Vector Machines.

3.3.1 Linear Discriminant Analysis

Linear Discriminant Analysis is a classification method that attempts to find the linear transformation of feature vectors into a feature space that produces the greatest distinction between classes, and then attempts use those to classify new samples based a hyperplane through that feature space (Haeb-Umbach, 1992). It achieves this by modeling the class conditional distribution of data $P(X|y = k)$ for each sentiment class $y = k$ where X represents the feature space. Probabilities of sentiment classifications are calculated using Bayes' rule. This feature distribution is modeled as a multivariate Gaussian distribution, with each class assumed to have the same covariance matrix, resulting in linear hyperplane decision surfaces.

This was not expected to perform very highly due to the simplicity of its method, but is a good baseline with which to compare the accuracy of other methods.

3.3.2 Logistic Regression

Logistic Regression is a model that uses the logistic function to attempt to calculate classification probabilities as the log-odds; a function of the predictor features, in order to apply linear regression to a classification problem. This allows for a less linear decision boundary by allowing for more independent feature relationships, as well as better predictions from discrete features (Prabhat, 2017). However it is also a relatively primative method, and is used along here alongside Linear Discriminant Analysis as a baseline comparison.

3.3.3 Random Forest Classifier

Random Forest Classification is an ensemble method that uses a large number of decision trees which each process the input feature data and attempt to independently determine a classification. A decision tree is a structured model of decision nodes that compute attributes and feed into further nodes until a conclusion node is reached, which outputs a classification. An example can be seen in Figure 3.3.1 . Decision trees are randomly constructed and trained around parts of the feature set, using Gini impurity, the likelihood of a new sample being incorrectly classified by the decision tree, as a measure of decision tree quality. These classifications are then compiled, and the most strongly supported classification is selected. This method aims to resolve the individual misclassifications any decision tree can have by having the final decision be the consensus of multiple trees, insulating the system from individual error (Gupte, 2014). Random Forest is a very accurate classification method for large datasets (Al Amrani, 2018).

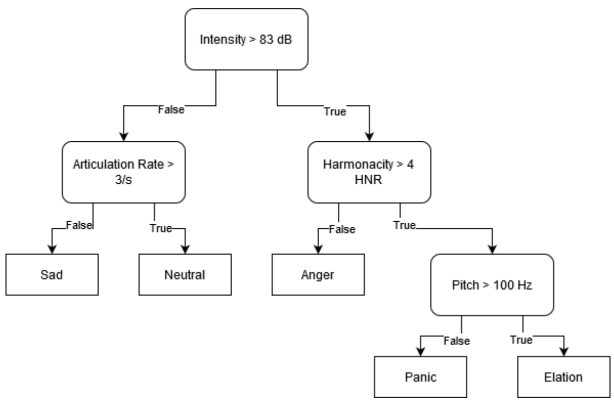

Figure 3.3.1. Example of a possible decision tree for sentiment analysis

As the construction properties of these decision trees can greatly influence the end result of the network, maximum tree depth and maximum features for a node split were varied on via grid search produce the highest accuracy for this data set.

3.3.4 Support Vector Machines

Support Vector Machines are a machine learning model that attempts to find the hyperplane that maximizes the margin between points of difference classes in the feature space, using points close to the hyperplane decision boundary (support vectors) to guide the position and orientation of the hyperplane to maximize the margin between classes (Al Amrani, 2018). Support Vector Machines map the input points to a new feature space by a non-linear transformation in which the hyperplane decision boundary is found. In this study the radial basis functin kernal was used, which transform the inputs by $K(x, x') = \exp(-\frac{\|x-x'\|^2}{2\sigma^2})$ where x and x' are pairs of inputs. It is considered one of the best classification methods for sentiment classification (Xia, 2011).

3.3.5 Layered Decision Tree of Trained Models

Some studies have shown increased classifier performance by incorporating multiple trained models in a decision tree based on group sentiment classifications (Sun, 2019). These train a top-level model to classify utterances into broader sentiment classes based on super-categories of sentiments that are easier to distinguish. Based on these classifications, the decision tree assigns them to additional models trained on limited data to more specifically distinguish between the sentiment classes of those super-categories.

3.3.6 Implementation of Models

Each of these were trained on folds of 90% of the features extracted from the MASC 's 20400 utterance dataset, and verified with the remaining 10% fold. Each of these folds's features were then standardized to their z-scores within their fold. Models were implemented using the Scikit-learn library in python using the train_test_split, LinearDiscriminantAnalysis, LogisticRegression, RandomForestClassifier and SVM procedures.

4
Results

4.1 Linear Discriminant Analysis, Logistic Regression, Random Forest Classifier and Support Vector Machines

The results of testing the first four models: Linear Discriminant Analysis, Logistic Regression, Random Forest Classifier and Support Vector Machines can be seen in Table 4.1.1. Of these four, Support Vector Machines achieved the highest accuracy on the test dataset, with approximately 53% accuracy for the five sentiment classifications. However, the significant difference in training and test accuracy for Support Vector Machines and in particular Random Forest Classifier is indicative of over-fitting to the training dataset, which could have negatively impacted their test accuracy.

Method	Train Accuracy	Test Accuracy
Linear Discriminant Analysis	0.4675	0.4480
Logistic Regression	0.4672	0.4510
Random Forest Classifier	0.9914	0.4701
Support Vector Machines	0.6148	0.5299

Table 4.1.1. Classification accuracy of different methods on both training and testing folds

4.2 Layered Decision Tree of Support Vector Machines

Due to Support Vector Machines having the highest accuracy among those four models, they were selected to be used to construct the layered decision tree of trained models. First a confusion tree was constructed using the Support Vector Machine model trained on the entire dataset, as shown in Table 4.2.1.

Sentiment	Anger	Elation	Neutral	Panic	Sadness
Anger	2494 (68%)	508 (14%)	166(5%)	418 (11%)	85 (2%)
Elation	949 (26%)	1744 (48%)	291 (8%)	494 (13%)	184 (5%)
Neutral	106 (3%)	210 (6%)	2626 (71%)	167 (5%)	573 (16%)
Panic	631 (17%)	453 (12%)	346 (9%)	1988 (54%)	256 (7%)
Sadness	107 (3%)	170 (5%)	759 (21%)	200 (6%)	2435 (66%)

Table 4.2.1. Confusion Matrix of Training data on Support Vector Machines

As indicated by Table 4.2.1, Anger, Elation and Panic form a distinct group of easily confusion sentiments by the Support Vector Machine, with Neutral and Sadness forming a separate distinct group of sentiments confused for the other. These two group are, however, more easily discriminated from from each other, with on average less than 20% confusion between sentiments of each group. From this a Decision Tree of Support Vector Machine can be constructed, with a top-level Support Vector Machine to identify the difference between these two distinct confusion groups, and then provide them to two separate Support Machines trained to better distinguish the sentiments within each confusion group from each other, as outlined in Figure 4.0.1 .

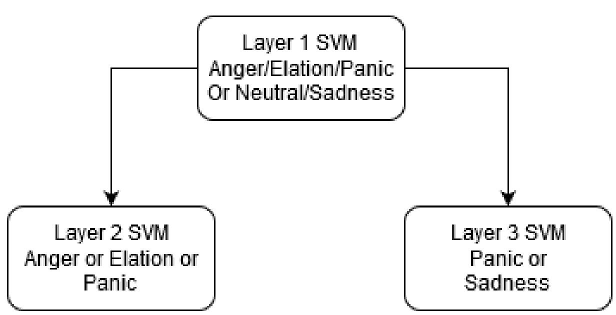

Figure 4.2.1. Diagram of Decision Tree of Support Vector Machines

Method	Train Accuracy	Test Accuracy
Support Vector Machines	0.6148	0.5299
Decision Tree Layer 1 SVM	0.8818	0.8701
Decision Tree Layer 2 SVM	0.6355	0.4789
Decision Tree Layer 3 SVM	0.7869	0.6121
Decision Tree Layer SVM Total	0.6961	0.5316

Table 4.2.2. Classification accuracy base SVM vs Decision Tree Layer SVMs and total Decision Tree Accuracy

The results of the Layered Decision Tree of Support Vector Machines are shown in Table 4.2.2, comparing the accuracy of individual layer's Support Vector Machine Models as well as the final accuracy of the total model against the original Support Vector Machine model.

5
Discussion

Overall, accuracy for these models given low complex feature sets and tested on Mandarin Chinese was low compared to the accuracy reported by many state of the art sentiment analysis models, achieving at best approximately 53% accuracy for five sentiment classes, compared to the 75-85% accuracy some studies have achieved for five sentiment classes (Majumder, 2018; Zhong, 2012; Sun, 2019). However, compared to the 60% accuracy achieved by more complex models tested on Mandarin Chinese (Chen, 2018), this accuracy is not entirely unexpected, especially given its low feature count.

What is unexpected however is the relatively small difference in accuracy between the primitive classification models used as a baseline such as Linear Discriminate Analysis and Logistic Regression, and the more complex classification models such as Random Forest Decision Trees and Support Vector Machines. Previous research has generally shown a dramatic difference in accuracy between these models (Prabhat, 2017). Similar results have been achieved by other studies comparing low complexity baseline models to higher complexity models for Mandarin Chinese datasets (Yang, 2008). Furthermore, the approximately 7% difference in accuracy between what was achieved by this study and the 60% achieved by a more complex model incorporating a high feature set is less than one would expect for a comparison of a low feature set model to a high feature set model, which can produce increases of 10-20% in classification accuracy (Chen, 2018; Gu, 2015). These findings are indicative that conventional methods of increasing model accuracy may not be as effective in languages outside of the most commonly tested datasets. This conclusion is supported by the unusual confusion matrix grouping of Neutral and Sadness, with low confusion between Sadness and Panic, which contradicts the usual findings regarding confusion between these sentiments in English (Sun, 2019). This is very likely a result of how these sentiments are expressed differently in Mandarin Chinese speech compared to more well studied languages for sentiment analysis. Additionally, the similar results of the Support Vector Machines compared to the Layered Decision Tree of Support Vector Machines do indicate that despite having what would conventionally be a core set of vital features, the feature set used was insufficient to allow a layered decision tree of trained models approach to be effective. These findings support the idea that features conventionally considered to have strong correlations to sentiment may not have those same correlations in other languages, and features not commonly considered important may be needed.

This study does acknowledge that the differences in training and test accuracy for the Random Forest Classifier and Support Vector Machines indicate a degree of over-fitting to the training data, due to a lack of validation, that likely negatively impacted their test accuracy. Validation to control for over-fitting should be applied in the future to correct for this error.

6 Conclusion

Sentiment analysis as a field has achieved impressive accuracy using increasingly complex and advanced methods tested on a few well studied language datasets. However, outside of those well studied languages, there is often a significant decrease in accuracy even with state of the art methods. In this study we have tested a number of well studied methods on a Mandarin Chinese speech dataset, and found that methods used to improve accuracy for well studied languages may not be as applicable outside of those languages.

This exposes a need for more careful future study of differences between languages in the context of sentiment analysis, and a reexamination of commonly held beliefs regarding approaching improving accuracy of sentiment analysis models. In the future we plan to further investigate alternative approaches to sentiment analysis to improve accuracy for Mandarin Chinese. Further analysis of what specific features performed well or poorly on the Mandarin Chinese dataset in this study, as well as exploration of features outside of the traditionally held essential for sentiment analysis is worth investigation. As well as a confirming accuracy for Random Forest Classifiers and Support Vector Machines with proper validation for over-fitting.

Bibliography

Al Amrani, Yassine; Mohamed, Lazaar; El Kadiri. 2018. *Random forest and support vector machine based hybrid approach to sentiment analysis*, Procedia Computer Science **127**, 511-520.

Ayadi, El; Kamel, Mohamed S.; Karray. 2011. *Survey on speech emotion recognition: Features, classification schemes,and databases*, Pattern Recognition **44(3)**, 572-587.

Chen, Feiyang; Luo, Ziqian. 2018. *Learning Robust Heterogeneous Signal Features from Parallel Neural Network for Audio Sentiment Analysis*, arXiv preprint arXiv:1811.08065.

Chen, Lijiang; Mao, Xia; Xue. 2012. *Speech emotion recognition: Features and classification models*, Digital Signal Processing **22(6)**, 1154-1160.

Gu, Yu; Postma, Eric; Lin. 2015. *Vocal emotion recognition with log-gabor filters*, AVEC 2015 - Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, co-Located with MM 2015, 25-31.

Gupte, Amit; Joshi, Sourabh; Gadgul. 2014. *Comparative study of classification algorithms used in sentiment analysis*, International Journal of Computer Science and Information Technologies **5(5)**, 6261-6264.

Haeb-Umbach, Reinhold; Ney, Hermann. 1992. *Linear discriminant analysis for improved large vocabulary continuous speech recognition*, 1992 International Conference on Acoustics, Speech, and Signal Processing (ICASSP) **1**, 13-16.

Li, Linchuan; Wu, Zhiyong; Wu. 2016. *Combining CNN and BLSTM to Extract Textual and Acoustic Features for Rec- ognizing Stances in Mandarin Ideological Debate Competition*, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH **08-12-Sept**, 1392-1396.

Liu, Dan; Lu, Lie; Zhang. 2003. *Automatic Mood Detection from Acoustic Music Data*, Proc. 4th International Society for Music Information Retrieval Conference.

Majumder, N; Hazarika, D; Gelbukh. 2018. *Multimodal sentiment analysis using hierarchical fusion with context modeling*, Knowledge-Based Systems **161**, 124-133.

Poria, Soujanya; Cambria, Erik; Howard. 2016. *Fusing audio, visual and textual clues for sentiment analysisfrom multimodal content*, Neurocomputing **174**, 50-59.

Prabhat, Anjuman; Khullar, Vikas. 2017. *Sentiment classification on big data using Naïve Bayes and logistic regression*, 2017 International Conference on Computer Communication and Informatics (ICCCI), 1-5.

Sun, Linhui; Fu, Sheng; Wang. 2019. *Decision tree SVM model with Fisher feature selection for speech emotion recognition*, Eurasip Journal on Audio, Speech, and Music Processing **2019**, 2.

Tao, Jianhua, Tan. 2005. *Affective Computing: A Review*, Affective Computing and Intelligent Interaction, 981-995.

Wang, Kunxia; An, Ning; Li. 2015. *Speech Emotion Recognition Using Fourier Parameters*, IEEE Transactions on Affective Computing **6**, 69-75.

Wu, Tian; Yang, Yingchun; Wu. 2006. *MASC: A Speech Corpus in Mandarin for Emotion Analysis and Affective Speaker Recognition*, IEEE Odyssey 2006: Workshop on Speaker and Language Recognition, 1-5.

Xia, Rui; Zong, Chengqing; Li. 2011. *Ensemble of feature sets and classification algorithms for sentiment classification*, Information sciences **181(6)**, 1138-1152.

Yang, Yingchun; Wu, Tian; Lv. 2008. *Simplified Deformation Compresation For Emotional Speaker Recognition*, 2008 6th International Symposium on Chinese Spoken Language Processing, 1-4.

Zhong, Jiang; Cheng, Yifeng; Yang. 2012. *Music Sentiment Classification Integrating Audio with Lyrics*, Journal of Information and Computational Science **9**, 35-44.