When Siri Knows How You Feel: Study of Machine Learning in Automatic Sentiment Recognition from Human Speech

Zhang Liu
Anglo-Chinese Junior College
Singapore
ifisay.617@gmail.com

Ng EYK
College of Engineering
Nanyang Technological University (NTU)
Singapore

***Abstract*— Opinions and sentiments are essential to human activities and have a wide variety of applications. As many decision makers turn to social media due to large volume of opinion data available, efficient and accurate sentiment analysis is necessary to extract those data. Hence, text sentiment analysis has recently become a popular field and has attracted many researchers. However, extracting sentiments from audio speech remains a challenge. This project explored the possibility of applying supervised Machine Learning in recognizing sentiments in English utterances on a sentence level. In addition, the project also aimed to examine the effect of combining acoustic and linguistic features on classification accuracy. Six audio tracks were randomly selected to be training data from 40 YouTube videos (monologue) with strong presence of sentiments. Speakers expressed sentiments towards products, films, or political events. These sentiments were manually labelled as negative and positive based on independent judgement of 3 experimenters. A wide range of acoustic and linguistic features were then analyzed and extracted using sound editing and text mining tools respectively. A novel approach was proposed, which used a simplified sentiment score to integrate linguistic features and estimate sentiment valence. This approach improved negation analysis and hence increased overall accuracy. Results showed that when both linguistic and acoustic features were used, accuracy of sentiment recognition improved significantly, and that excellent prediction was achieved when the four classifiers were trained respectively, namely kNN, SVM, Neural Network, and Naïve Bayes. Possible sources of error and inherent challenges of audio sentiment analysis were discussed to provide potential directions for future research.**

Keywords – Sentiment Analysis; Natural Language Processing; Machine Learning; Affective Computing; Data Analytics; Speech Processing; Computational Linguistic.

I. INTRODUCTION

Sentiment analysis is the field of study that analyses opinions, sentiments, appraisals, attitudes, and emotions toward entities and their attributes [1]. Opinions and sentiments are essential to human activities and have a wide variety of applications. As many decision makers turn to social media due to large volume of opinion data available, efficient and accurate sentiment analysis is necessary to extract those data. Business organizations in different sectors use social media to find out consumer opinions to improve their products and services. Political party leaders need to know the current public sentiment to come up with campaign strategies. Government agencies also monitor citizens' opinions on social media. Police agencies, for example, detect criminal intents and cyber threats by analyzing sentiment valence in social media posts. In addition, sentiment information can be used to make predictions, such as in stock market, electoral politics and even box office revenue. Moreover, sentiment analysis that moves towards achieving emotion recognition can potentially enhance psychiatric treatment as emotions of patients are more accurately identified.

Since 2000, researchers have made many successful attempts in text sentiment analysis. In comparison, audio sentiment analysis does not seem to receive as much attention. It is, however, equally significant as text sentiment analysis. Many people in the contemporary society share their opinions using online-based multimedia platforms such as YouTube videos, Instagram stories, TV talk shows and TED talks. It is difficult to manually classify sentiments in them due to the sheer amount of data. With the help of machine automation, we can recognize, with an acceptable accuracy, the general sentiments about certain products, movies, and socio-political events, hence aiding decision-making process of corporations, societal organizations and governments.

This project explored the possibility of using a machine learning approach to recognize sentiments accurately and automatically from natural audio speech in English. In addition, the project also aimed to examine the effect of combining acoustic and linguistic features on classification accuracy. Training data consisted of 150 speech segments extracted from 6 YouTube videos of different genres. Both acoustic features and linguistic features was examined in order to increase the accuracy of automatic sentiment recognition. Sentiments were categorized into 2 target classes, positive and negative.

II. Literature Review

There were previous attempts to combine acoustic and linguistic features of speech in sentiment analysis. Chul & Narayanan (2005) [2] explored the detection of domain-specific emotions (negative and non-negative) using language and discourse information in conjunction with acoustic correlates of emotion in speech signals. The database consists of spoken speech obtained from a call center application. using. Their results showed that combining all the information, rather than using only acoustic information, improved emotion classification by 40.7% for males and 36.4% for females. This study suggested a comprehensive range of features and provided some insights for my project: acoustic features (Fundamental Frequency (F0), Energy, Duration, Formants), and textual features (emotional salience, discourse information). However, with its speech data collected from a call center, the research focused on emotions in human-machine interactions, rather than in natural human speech.

Another research, Kaushik & Sangwan & Hansen (2013) [3], provided an alternative source of speech data - YouTube videos. In this study, the authors proposed a system for automatic sentiment detection in natural audio streams on social media platform such as YouTube. The proposed technique used Part of Speech (POS) tagging and Maximum Entropy modeling (ME) to design a text-based sentiment detection model. Using decoded Automatic Speech Recognition (ASR) transcripts and the ME sentiment model, the proposed system was able to estimate sentiments in YouTube videos. Their results showed that it was possible to perform sentiment analysis on natural spontaneous speech data despite poor word error rates. This study provided a systematic approach and proved that audio sentiment analysis is possible. It did not, however, include enough acoustic features of audio speech, possibly due to the limitation of document-level analysis.

Ding, *et al.* proposed a holistic lexicon-based approach [4] to solve the problem of insufficient acoustic features by exploiting external evidences and linguistic conventions of natural language expressions. Inspired by above work, a simplified sentiment score model was proposed in this project. The model was useful in sentence level audio speech analysis. The detail of the method will be explained in section III.

III. Methodology

Fig. 1. An overview of the methodology of this project.

3.1 DATABASE

The speech data used in the experiments were obtained from YouTube, a social media platform. This source was chosen because thousands of YouTube users share their personal opinions or reviews on their channels. Hence, there is a huge amount of accessible speech data containing sentiment valence. More importantly, their ways of speaking are usually closest to natural, spontaneous human speech. Six videos were randomly selected from 40 YouTube videos that had strong presence of negative or positive sentiments. Subject matters included: 1) Product Review; 2) Movie Review; 3) Political Opinion.

During the pre-processing stage, the videos were converted into '.wav' files. Speech transcriptions were generated using the Automatic Speech Recognition (ASR) software, Speechmatics (https://www.speechmatics.com) and checked manually to increase reliability. Each sound file (.wav) was then edited in the vocal toolkit, Praat (http://www.fon.hum.uva.nl/praat/), to match the transcriptions to corresponding sound segments. The TextGrid annotation (as shown in Figure 2) included 2 tiers, transcription text and numbering, which were useful in keeping track of the data. Meanwhile, the sound file was segmented into smaller sections containing 1 to 5 sentences of relevant meaning and the same sentiment. Each segment was pre-assigned a sentiment label ('negative' or 'positive') based on independent judgement of 3 experimenters so as to minimize bias and subjective errors. There was a total of 150 sound segments (including 70 positive, 80 negative) in the data set. The segmentation process was necessary as most opinion videos contain mixed sentiments.

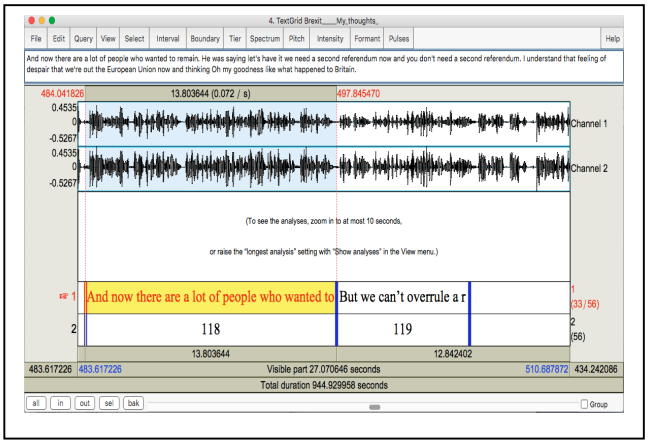

Fig. 2. Using Praat to annotate speech.

3.2 FEATURE EXTRACTION

3.2.1 LINGUISTIC FEATURES

Natural Language Processing toolkit, Orange 3-Text Mining, was used in this stage. Speech transcripts were transformed into lowercase, tokenized into words, and normalized using WordNet Lemmatizer. Part of Speech (POS) tagger was used to label each word as, for instance, a noun, a verb or an adjective, in order to preserve the linguistic function of each word in the sentence. The workflow and text processor parameters were shown in Figure 3 and 4.

Textual feature extraction was done by filtering the emotionally salient words (negatively connoted words and positively connoted words). Words in the training corpus were looked up against Harvard General Inquirer and Opinion Lexicon by Bing, Liu [5] to decide if they were negatively or positively connoted. The frequencies of negatively and positively connoted words in each segment were then counted respectively and the numerical values were stored in the training data set.

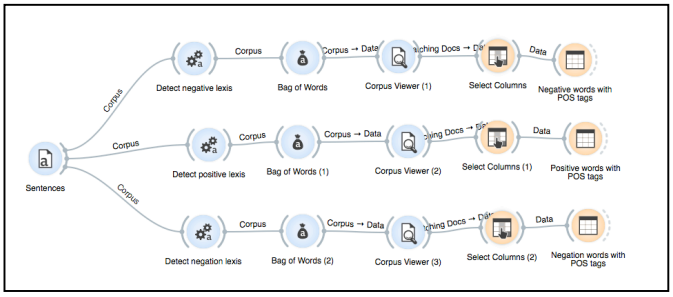

Fig. 3. Orange 3 Text Mining workflow.

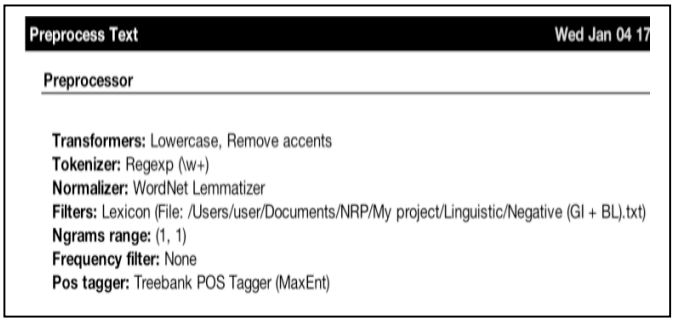

Fig. 4. Text preprocessor parameters.

For negation cues, a similar approach was adopted to look up words against a list of explicit negation cues (compiled manually) as shown in Table I.

TABLE I. LIST OF NEGATION CUES

aint	doesnt	havent	lacks	nobody	prevent
arent	dont	havnt	mightnt	none	rarely
barely	doubt	improbable	mustnt	nor	scarcely
cannot	few	isnt	neednt	not	seldom
cant	hadnt	lack	neither	nothing	shant
darent	hardly	lacked	never	nowhere	shouldnt
didnt	hasnt	lacking	no	oughtnt	unlikely
wasnt	werent	without	wouldnt	little	

Last but not least, a simplified "Sentiment Score" model was proposed to "integrate" all the linguistic features that provide emotion-related information. The Sentiment Score reflected the sentiment of an opinion, with sentiment defined as the valence. For every opinion segment O, a

sentiment score, $f(O)$, was calculated using the formula below:

$$f(O) = pos(O) - neg(O) + 2 \times (\sum(-1)^{neg\_pos(w)}) - \sum(-1)^{neg\_neg(w)}),$$

where $pos(O)$ is the number of positive words in the opinion segment, O; $neg(O)$ is the number of negative words in the opinion segment, O; $neg\_pos(w)$ is the number of times each positive word is negated, and similarly, $neg\_neg(w)$ is the sum of the number of times each negative word is negated. Note that $\sum(-1)^{neg\_pos(w)})$ and $\sum(-1)^{neg\_neg(w)}$ were counted manually to give the most reliable values. The following are some advantages of this model.

1) The problem of multiple negation can be solved. When a word is negated twice (with our loss of generality, suppose it is a positive word, as in "can't live without"), the formula will correctly give a positive value that signifies positive sentiment.
2) It allows semi-automation negation analysis and has potential to be developed into a fully automated process.

3.2.2 ACOUSTIC FEATURES

Acoustic features of the sound segments were extracted manually using built-in functions in Praat, as shown in Figure 5. In order to achieve a more comprehensive representation of the sound, I chose a sufficiently wide range of acoustic features: intensity (amplitude, total energy, mean power), pitch (maximum pitch, average pitch, standard deviation, mean absolute slope), and voice quality (jitter, shimmer, Mean harmonics-to-noise ratio). Considering the inherent differences in pitch between females and males, an attribute "gender" was included to normalize the data.

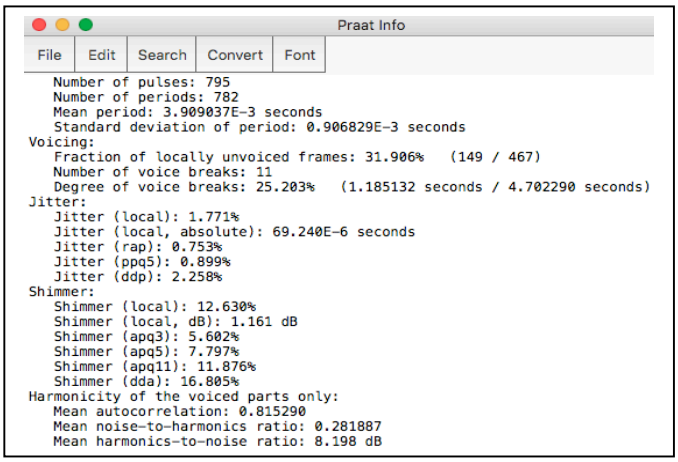

Fig. 5. Extracting acoustic features using Praat.

3.3 MACHINE LEARNING

In the Orange Canvas Software [6], kNN, NN, Naïve Bayes and SVM were used to evaluate the proposed method. Extracted features were sent to appropriate classifier (Figure 6). Sentiment label (positive, negative) was selected as the target class and the rest of the features as attributes. Stratified 10-folds cross-validation method was used to measure model performance. Hence, each time the dataset was split into ten folds and one out of ten folds was randomly selected for testing. After multiple experiments, optimal configuration for each classifier was determined and used in the machine learning process.

TABLE II. OPTIMAL CONFIGURATION FOR DIFFERENT CLASSIFIERS

Classifier	Optimal Configurations
k Nearest Neighbors	▪ k = 74 (weighting by distances) ▪ Euclidean (normalize continuous attributes)
Naïve Bayes	▪ Prior: Relative Frequency ▪ Conditional: M-Estimate (parameter = 2.0) ▪ Size of LOESS window = 1.0 ▪ LOESS sample points = 11 ▪ Adjust threshold
Neural Network	▪ Hidden layer neurons = 11 ▪ Regularization factor = 1.0 ▪ Max iterations = 300 ▪ Normalize data
Support Vector Machine	▪ C-SVM (C = 1.00) ▪ Linear Kernel, x- y ▪ Numerical tolerance = 0.0010 ▪ Estimate class probabilities ▪ Normalize data

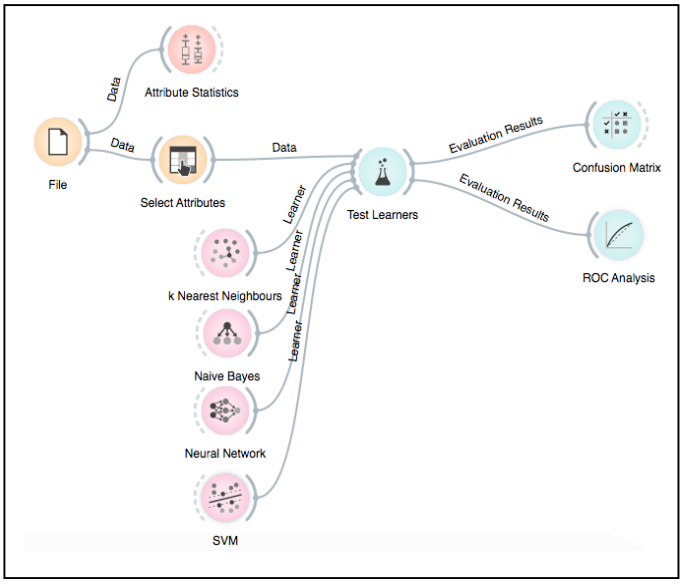

Fig. 6. Illustration for machine learning workflow.

IV. Results And Discussions

4.1 RESULTS ANALYSIS

The evaluation will be focused on Area Under the ROC Curve (AUC) as it has "better statistical foundations than most other measures" [7] ROC Area Benchmark: 1.0: perfect prediction; 0.9: excellent prediction; 0.8: good prediction; 0.7: mediocre prediction; 0.6: poor prediction; 0.5: random prediction; <0.5: something wrong. [7] As shown in Table II, accuracy improved significantly when both acoustic and linguistic features were used, instead of only acoustic features or only linguistic features. When both acoustic and linguistic features were extracted, excellent classification of sentiments was achieved when the four classifiers were trained, with kNN, SVM and Neural Network having higher accuracies. The shapes of ROC curves for these four classifiers resembled the shape of ROC curve for excellent prediction (Figure 7 and Figure 8 [8]).

TABLE III. AUC WHEN DIFFERENT CLASSIFIERS & FEATURES ARE USED

Classifier	AUC (acoustic features only)	AUC (linguistic features only)	AUC (Both acoustic features & linguistic features)
kNN	0.8750	0.8420	0.9321 > 0.9
Naïve Bayes	0.7964	0.8348	0.8929 ≈ 0.9
Neural Network	0.9018	0.8384	0.9304 > 0.9
SVM	0.8589	0.8607	0.9429 > 0.9

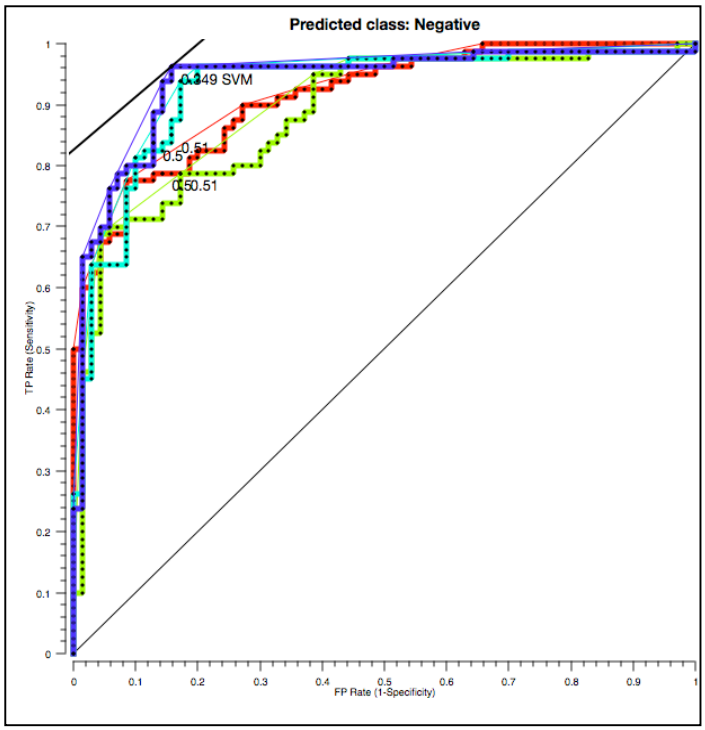

Fig. 7. ROC curves for different classifiers when both acoustic and linguistc features were used.

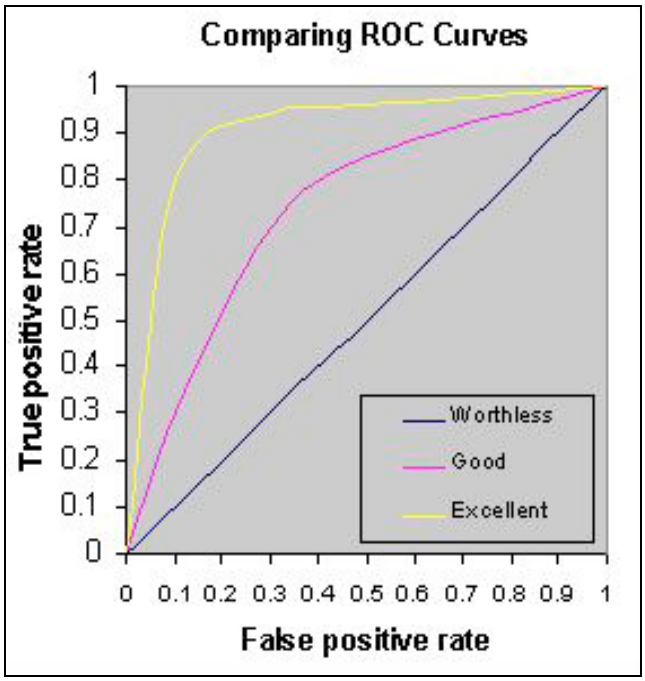

Fig. 8. Shapes of ROC curves indicate different levels of accuracy.

4.2 LIMITATIONS AND SOURCES OF ERROR

The speech corpus might not be large enough. There has to be sufficient representation of the different sentiments and different speech types in order for more comprehensive learning, and hence more accurate recognition. A standard sentiment analysis database could be used as a benchmark to compare the accuracy of this linguistic/acoustic model against other models. The average Word Error Rate of the Automated Speech Recognition (ASR) software Speechmatics is reported to be 33%. Hence, transcription errors might still be present even after manual check, affecting the accuracy of linguistic analysis. Negation analysis is subject to human error and there might be inaccurate detection of context-specific meanings of polysemes[1].

4.3 INHERENT CHALLENGES

Sentiment analysis is a challenging task due to ambiguities in language, such as subtlety, concession, manipulation, sarcasm and ironies in speech. To address this problem, an accurate conclusion might entail examination of other features such as physiological symptoms (blood pressure etc.) and facial expressions. Although inaccuracies arising from ambiguities could be minimized by analyzing data from multiple dimensions,

[1] A word or lexical unit that has several or multiple meanings

cultural differences and multilinguality[2] further complicate the process. Due to differences in cultural backgrounds, the ways people express their sentiments vary among individuals. (For example, the way a Japanese expresses a sentiment differs from the way an American expresses the same sentiment). Moreover, sentiment expressions depend on contexts of speech, and hence vary even for the same person at different times. In addition, the speakers might constantly change subject or compare with another subject, which might be hard to detect.

There are also issues with mutual interpretability. Interjections that express feelings (such as "urggghhh") might be deemed as irrelevant by the machine. It might be hard, if not impossible, for the machine to "master" contextual knowledge such as some exophoric references[3] to historical figures ("the German dictator", which refers to Hitler). The issue becomes more significant when dialects are used. For example, the negation analysis is based on Standard English usage, which might not be useful for other varieties of English. Speakers of certain dialects like African American Vernacular English (AAVE) usually employ double negatives to emphasize the negative meaning.

V. Conclusion And Future Work

In this study, we have built a machine learning model combining acoustic and linguistic features. As the results have shown, this model has significantly higher accuracy than models with only acoustic or only linguistic features. Under this model, excellent prediction can be achieved. Although limitations and challenges are real and a considerable amount of manual work is necessary, the positive results of this study have clearly suggested to the possibility of achieving a fully automated audio sentiment analysis in future.

Based on the limitations and challenges discussed in section IV, the following three main directions of research are proposed.

1. From *audio* sentiment analysis towards *video* sentiment analysis by incorporating facial expression features, and further towards *multi-dimensional* sentiment analysis by incorporating physiological features such as blood pressure and heart rate.
2. From *semi*-automatic sentiment analysis towards *fully* automatic sentiment analysis, by reducing the amount of manual processing of data.
3. From *sentiment* recognition towards *emotion* recognition, by enabling classification of specific emotions such as fear, anger, happiness, sadness.

Artificial Intelligence (AI) is becoming an increasingly interdisciplinary field. To achieve the above research goals, cross-discipline cooperation is crucial. Solutions to the challenges of language/emotion recognition and understanding can be inspired by diverse fields from mathematics and sciences, which provide us with quantitative methods and computational models, to humanities and fine arts, which shed light on qualitative analysis and feature selection. From a neuroscience perspective, learning about how the human brain perceives and processes sentiments and emotions might inspire a better machine learning architecture for sentiment prediction. Mathematical modelling could be useful as well: the high complexity of emotions should be captured more comprehensively by mapping the emotion of each utterance in multi-dimensional vector space. Linguistics theories also imply that language is meaningless without context (the socio-cultural background of the speaker, the conversation setting, and the general mood). It is a timely reminder for Natural Language Processing (NLP) researchers to go beyond *content analysis* – dissecting language as an isolated entity only made up of different parts of speech – and aim for "*context analysis*". Without being context-aware, AI will only be machines with "high Intelligence Quotient (IQ)" but "low emotional intelligence quotient (EQ)". Emotion theory in drama and acting also provides some insights for developing affective, sentient AI. For example, emotions can be conveyed through subtle means such as silence, cadence, and paralinguistic features (kinesics, i.e. body language and proxemics i.e. use of space etc.). This will give us directions in selecting and extracting features salient to sentiment.

[2] Multilinguality is a characteristic of tasks that involve the use of more than one natural language. (Kay, n.d.)

[3] *Exophoric reference is* referring to a situation or entities outside the text. (University of Pennsylvania, 2006)

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my project supervisor Professor Eddie Ng for being open-minded about my project topic, without which I could not have been able to delve deep into my field of interest. His insightful suggestions and unwavering support has guided me through doubts and difficulties.

REFERENCES

[1] Liu, Bing. (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press.

[2] Chul Min Lee & Shrikanth S. Narayanan. (2005). Toward Detecting Emotions in Spoken Dialogs. *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 13, NO. 2, MARCH 2005.*

[3] Kaushik, Lakshmish & Sangwan, Abhijeet & Hansen, John H.L. (2013). A Holistic Lexicon-Based Approach to Opinion Mining. *IEEE.*

[4] Ding, Xiaowen, Bing Liu, and Philip S. Yu. A Holistic Lexicon-Based Approach to Opinion Mining. In *Proceedings of the Conference on Web Search and Web Data Mining (WSDM- 2008)*. 2008.

[5] Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA

[6] Demšar, J., Curk, T., & Erjavec, A. (2013). Orange: Data mining toolbox in Python. *Journal of Machine Learning Research, 14*, 2349−2353.

[7] Unknown. Cornell University. (2003). Retrieved from: https://www.cs.cornell.edu/courses/cs578/2003fa/performance_measures.pdf

[8] Tape, Thomas G. (n.d.). *Interpreting Diagnostic Tests*. University of Nebraska Medical Center. Retrieved from: http://gim.unmc.edu/dxtests/roc3.htm