25th International Joint Conference on Artificial Intelligence

New York City, USA, 9th-15th July, 2016

Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016)

New York City, USA, 10th July, 2016

SAAIP 2016

Volume Editors

Sivaji Bandyopadhyay
Department of Computer Science and Engineering,
Jadavpur University, Kolkata - 700032, India.
E-mail: sivaji_cse_ju@yahoo.com

Dipankar Das
Department of Computer Science and Engineering,
Jadavpur University, Kolkata - 700032, India.
E-mail: dipankar.dipnil2005@gmail.com

Erik Cambria
School of Computer Engineering,
Nanyang Technological University, Singapore 639798.
E-mail: cambria@ntu.edu.sg

Braja Gopal Patra
Department of Computer Science and Engineering,
Jadavpur University, Kolkata - 700032, India.
E-mail: brajagopal.cse@gmail.com

Publisher

CEUR Workshop Proceedings
http://ceur-ws.org/Vol-1619
ISSN 1613-0073

Preface

In recent times, research activities in the areas of Opinion, Sentiment, Emotion and/or Mood in natural language texts, speech, music and other media have become the mainstream research under the umbrella of subjectivity analysis and affective computing. These tasks are considered vital from various academic and commercial perspectives since a decade. The popularity of the Internet and the rapid expansion of social media have made available online a variety of user generated contents. However, the major challenges are how to process the user generated contents such as texts, audio and images and how to organize them in some meaningful ways.

The common interest areas where Artificial Intelligence (AI) meets sentiment analysis can be viewed from four aspects of the problem and the aspects can be grouped as Object identification, Feature extraction, Orientation classification and Integration. The existing reported solutions or available systems are still far from being perfect or fail to meet the satisfaction level of the end users. The main issue may be that there are many conceptual rules that govern sentiment and there are even more clues (possibly unlimited) that can convey these concepts from realization to verbalization of a human being. Human psychology may provide the unrevealed clues and govern the sentiment realization. The important issues that need attention include how various psychological phenomena can be explained in computational terms and which AI concepts and computational methodologies will be proved as the most useful ingredients from the psychologist's point of view.

Sentiment analysis from natural language texts is a multifaceted and multidisciplinary problem. Research efforts are being carried out for identification of positive or negative polarity of the evaluative text and also for the development of devices that recognize human affect, display and model emotions from textual contents. Identifying strength of sentiment in figurative texts or aspects and categories from the reviews, detecting stance from the tweet data, identifying the psychological condition of persons from chat even detecting sentiment in clinical texts and the moods from music etc. are the recent trends in the field of sentiment analysis.

Mood analysis from music is an emerging area in Music Information Retrieval (MIR). The popularity of downloading and purchasing of music from online music shops have increased. Similarly, with rapid evolvement of technology, music is just a few clicks away, on almost any personal gadget be it computers, portable music players, or smart phones. This fact underlines the importance of developing an automated process for its organization, management, search as well as generation of playlists and various other music related applications. Recently, MIR based on emotions or moods has attracted the researchers from all over the world because of its highly motivated implications in human computer interactions.

In addition to Question Answering or Information Retrieval systems, Topic-sentiment analysis is being applied as a new research method for mass opinion estimation (e.g., reliability, validity, sample bias), psychiatric treatment, corporate reputation measurement, political orientation categorization, stock market prediction, customer preference or public opinion study and so on. Techniques from Artificial Intelligence play the important roles in these tasks. In recent times, regular research papers continue to be published in reputed conferences like ACL, NAACL, EMNLP, COLING, IJCNLP, CICLing, EACL etc. The Sentiment Analysis Symposiums are also drawing the attention of the research communities from every nook of the world. There has been an increasing number of efforts in shared tasks such as SemEval 2007 Task: Affective Text, SemEval 2013-2016 Task: Sentiment Analysis on Twitter, SemEval 2014-2016: Aspect-Based Sentiment Analysis, SemEval 2015: Sentiment Analysis of Figurative Language in Twitter, SemEval 2015: CLIPEval Implicit Polarity of Events, SemEval 2016: Detecting Stance in Tweets, SemEval 2015-2016: Clinical TempEval, TAC 2008 Opinion Summarization

task, TREC-BLOG tracks since 2006, and relevant NTCIR tracks since 6th NTCIR that aimed to focus on different issues of opinion and emotion analysis. Research activities on Sentiment Analysis have been performed in several languages other than English. The shared task Sentiment Analysis in Indian Languages (SAIL) Tweets in 2015 has been organized to detect the sentiment from Bengali, Hindi and Telugu tweets. The shared task Aspect-Based Sentiment Analysis in SemEval 2016 has also targeted sentiment analysis in the languages like Arabic, Chinese, Dutch, French, Russian, Spanish and Turkish including English. Some of the important names e.g., MediEval: Emotion in Music and MIREX: Audio Music Mood Classification are the evaluation campaigns for the mood classification from music using audio. The 16^{th} International Society for Music Information Retrieval (ISMIR) is one of the most reputed conferences in the field of music and has published many papers related to music mood.

Several communities from sentiment analysis have engaged themselves to conduct relevant conferences, e.g., 6^{th}Affective Computing and Intelligent Interfaces (ACII), 5th Annual Conference Behavioural Models & Sentiment Analysis Applied to Finance in 2015, symposiums such as Sentiment Analysis Symposium in 2015, and workshops such as Sentiment Analysis Innovation, "Sentiment and Subjectivity in Text" collocated with COLING-ACL 2006, "Sentiment Analysis – Emotion, Metaphor, Ontology and Terminology (EMOT)" in LREC 2008, Opinion Mining and Sentiment Analysis (WOMSA) 2009, "Topic-Sentiment Analysis for Mass Opinion Measurement (TSA)" in CIKM 2009, "Computational Approaches to Analysis and Generation of Emotion in Text" in NAACL 2010, 6^{th} Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA) in EMNLP 2015, FLAIRS 2011 special track on "Affect Computing", 5^{th}Sentiment Elicitation from Natural Text for Information Retrieval and Extraction (SENTIRE 2015), 5^{th} Workshop on EMOTION, SOCIAL SIGNAL, SENTIMENT AND LINKED OPEN DATA (ESLOD 2014) in the satellite of LREC 2014, "Practice and Theory of Opinion Mining and Sentiment Analysis" in conjunction with KONVENS-2012 (PATHOS-2012), Intelligent Approaches applied to Sentiment Mining and Emotion Analysis (WISMEA, 2012), Socio-Affective Computing, Emotion and Sentiment in Social and Expressive Media, workshop on "Issues of Sentiment Discovery and Opinion Mining (WISDOM, 2012)", and a bunch of special sessions like Sentiment Analysis for Asian Languages (SAAL, 2012), Brain Inspired Natural Language Processing (BINLP, 2012), Advances in Cognitive and Emotional Information Processing (ACEIP, 2012), Language Technologies for Indian Social Media in 2014 and so on.

Since our previous three workshops in conjunction with the International Joint Conference on NLP (IJCNLP) in Chiang Mai, Thailand during Nov. 7-13, 2011, International Conference on Computational Linguistics (COLING) in Mumbai, India during Dec. 8-15, 2012 and with the International Joint Conference on NLP (IJCNLP) in Nagoya, Japan during the period October 14-18, 2013 were quite successful (with 20, 14, and 10 submissions and more than 30 participants from many countries). Prof. Eduard Hovy and Prof. James Martin were the keynote speakers for the first and second versions of this workshop, respectively. Inspired by the objectives we aimed at in the first three editions of the workshop, the warm responses and feedbacks we received from the participants and attendees and the final outcome, the purpose of the 4^{th} edition of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016) in conjunction with International Joint Conference on Artificial Intelligence (IJCAI) 2016, is to create a framework for presenting and discussing the challenges related to sentiment, opinion, emotion, and mood analysis in the ground of NLP. This workshop also aims to bring together the researchers in multiple disciplines such as computer science, psychology, cognitive science, social science and many more who are interested in developing next generation machines that can recognize and respond to the sentimental states of the human users. This time we received

twelve submissions and finally eight papers have been accepted.

The keynote presented by Professor Björn Schuller is on seven essential principles to make multimodal sentiment analysis that works in the wild. A web-based interactive speech emotion classification system (WISE) that allows users to upload speech data and automatically classify the emotions was provided by Sefik Emre Eskimez, Melissa Sturge-Appley, Zhiyao Duan and Wendi Heinzelman.

Jasy Liew Suet Yan, Howard R. Turtle presented a work where a set of 48 emotion categories is discovered for the first time inductively from 5,553 annotated tweets through a small-scale content analysis by trained or expert annotators and then a set of 28 emotion categories was refined and tested to find out how representative they are on a larger set of 10,000 tweets obtained through crowd sourcing. While the most existing work in domain adaptation has focused on feature-based and/or instance-based adaptation methods, Bo Wang Maria Liakata, Arkaitz Zubiaga, Rob Procter and Eric Jensen set out to find an effective approach for tackling across-domain emotion classification task on a setof Twitter data involving social media discourse around arts and cultural experiences, in the context of museums. Determining an individual's personality traits is an important concept in Psychology that is also synchronized with our workshop's themes. In this regard, Edward P. Tighe, Jennifer C. Ureta, Bernard Andrei L. Pollo, Charibeth K. Cheng, and Remedios de Dios Bulos conducted research that aims to perform feature reduction techniques on linguistic features from essays and classify the author's personality traits based on the reduced feature set.

Sandeep Sricharan Mukku, Nurendra Choudhary and Radhika Mamidi explored various Machine Learning techniques for the classification of Telugu sentences into positive or negative polarities. In the domain of Bio medical Natural Language Processing (Bio-NLP), Anupam Mondal, Ranjan Satapathy, Dipankar Das and Sivaji Bandyopadhyay described a hybrid approach which is the combination of both linguistic and machine learning approaches to extract the contextual sense-based information from a medical corpus.

Rafal Rzepka, Kohei Matsumoto and Kenji Araki introduced a novel method for utilizing web mining and semantic categories for determining automatically if a given act is worth praising or not and reported how existing lexicons used in affective analysis and ethical judgment can be combined for generating useful queries for knowledge retrieval from a 5.5 billion word blog corpus. Finally, an approach to detect the sentiment of a song based on its multi-modality natures (text and audio) is presented by Harika Abburi, Eswar Sai Akhil Akkireddy, Suryakanth V Gangashetty and Radhika Mamidi.

We thank all the members of the Program Committee for their excellent and insightful reviews, the authors who submitted contributions for the workshop and the participants for making the workshop a success. We also express our thanks to the IJCAI 2016 Organizing Committee and Local Organizing Committee for their support and cooperation in organizing the workshop.

Organizing Committee
4thWorkshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016)
IJCAI 2016
July 10, 2016
New York, USA

Organizing Committee

Sivaji Bandyopadhyay	Jadavpur University, Kolkata (India) (Chair)
Dipankar Das	Jadavpur University, Kolkata (India) (Co-Chair)
Erik Cambria	Nanyang Technological University, Nanyang (Singapore)
Braja Gopal Patra	Jadavpur University, Kolkata (India)

Program Committee

Alexandra Balahur	European Commission Joint Research Centre (Italy)
Amitava Das	IIIT Sri City (India)
Alexander Gelbukh	Instituto Politécnico Nacional (Mexico)
Diana Inkpen	University of Ottawa (Canada)
Vladimir Ivanov	Kazan Federal University (Russia)
Saif Mohammad	NRC (Canada)
Vincent Ng	University of Texas at Dallas (USA)
Viviana Patti	University of Turin (Italy)
Veronica Perez-Rosas	University of Michigan (USA)
Paolo Rosso	Technical University of Valencia (Spain)
Patrick Saint-Dizier	IRIT-CNRS (France)
Yohei Seki	University of Tsukuba (Japan)
Swapna Somasundaran	Educational Testing Services (USA)
Veselin Stoyanov	Johns Hopkins University (USA)
Stan Szpakowicz	EECS, University of Ottawa (Canada)
Alessandro Valitutti	University of Helsinki (Ireland)
Michael Zock	CNRS-LIF (France)

Table of Contents

Program Schedule

Session 1:

11:00 - 11:10	Opening Remarks
11:10 - 12:10	**Invited Talk:** 7 Essential Principles to Make Multimodal Sentiment Analysis Work in the Wild *Björn W. Schuller*
12:10 - 12:30	WISE: Web-based Interactive Speech Emotion Classification *Sefik Emre Eskimez, Melissa Sturge-Apple, Zhiyao Duan and Wendi Heinzelman*
12:30 - 14:00	Lunch

Session 2:

14:00 - 14:20	Exposing a Set of Fine-Grained Emotion Categories from Tweets *Jasy Suet Yan Liew and Howard R. Turtle*
14:20 - 14:40	SMILES: Twitter Emotion Classification using Domain *Bo Wang, Maria Liakata, Arkaitz Zubiaga, Rob Procter and Eric Jensen*
14:40 - 15:00	Personality Trait Classification of Essays with the Application of Feature Reduction *Edward P. Tighe, Jennifer C. Ureta, Bernard Andrei L. Pollo, Charibeth K. Cheng and Remedios De Dios Bulos*
15:00 - 15:20	Enhanced Sentiment Classification of Telugu Text using ML Techniques *Sandeep Sricharan Mukku, Nurendra Choudhary and Radhika Mamidi*
15:30 - 16:00	Coffee

Program Schedule

Session 3:

16:00 - 16:20	A Hybrid approach based Sentiment extraction from Medical context *Anupam Mondal, Ranjan Satapathy, Dipankar Das and Sivaji Bandyopadhyay*
16:20 - 16:40	Praiseworthy Acts Recognition Using Web-based Knowledge and Semantic Categories *Rafal Rzepka, Kohei Matsumoto and Kenji Araki*
16:40 - 17:00	Multimodal Sentiment Analysis of Telugu Songs *Harika Abburi, Eswar Sai Akhil Akkireddy, Suryakanth Gangashetti and Radhika Mamidi*
17:00 - 17:30	Open Discussion and Valedictory

7 Essential Principles to Make Multimodal Sentiment Analysis Work in the Wild

Björn W. Schuller[1,2]
[1]Department of Computing, Imperial College London, United Kingdom
[2]Chair of Complex & Intelligent Systems, University of Passau, Germany
bjoern.schuller@imperial.ac.uk

Abstract

Sentiment Analysis is increasingly carried out on multimodal data such as videos taken in everyday environments. This requires robust processing across languages and cultures when aiming for mining of opinions from the 'many'. Here, seven key principles are laid out to ensure a high performance of an according automatic approach.

1 Introduction

Sentiment Analysis (SA) recently found its way beyond pure text analysis [Cambria *et al.*, 2015], as sentiment is increasingly expressed also via video 'micro blogs', short clips, or other forms. Such multimodal data is usually recorded 'in the wild' thus challenging today's automatic analysers. For example, one's video-posted opinion on a movie may contain scenes of this movie, requiring subject tracking, and music in the background may need to be overcome for speech recognition and voice analysis. Here, I provide 'essential principles' to make a multimodal SA work despite such challenges.

2 The Principles

Seven selected recommendations to make a multimodal SA system 'ready for the wild' are given with a short statement :

Make it Multimodal – But Truly. Multimodal SA is often carried out in a late fusion manner, as e. g., (spoken) language, acoustics, and video analysis operate on different time levels and monomodal analysers prevail. However, recent advances in synergistic fusion allow for further exploitation of heterogeneous information streams such as analysis of cross-modal behaviour-synchrony to reveal, e. g., regulation.

Make it Robust. Robustness is a obviously key handling real-world data. Effective denoising and dereverberation can these days be reached by data-driven approaches such as (hierarchical) deep learning. Beyond, recognition of occlusions, background noises, and alike should be used in the fusion to dynamically adjust weights given to the modalities

Train it on Big Data. A major bottleneck for SA beyond textual analysis is the lack of suited 'big' (ideally multimodal) training data. While data is usually 'out there' (such as videos on the net), it is the labels that lack. Recent cooperative learning approaches such as by dynamic active learning and semi-supervised learning combined with (gamified) crowdsourcing can help to efficiently build a large training corpus. Smart pre-selection of suited material from large resources can further improve efficiency.

Make it Adapt. It has repeatedly been shown that in-domain training improves SA [Cambria *et al.*, 2015]. Recent transfer learning provides a range of solutions to adapt to a new domain even if only little and/or unlabelled data should exist from it. Subject adaptation is another key aspect.

Make it Context-aware. Temporal context modelling can be learnt, e. g., by LSTM recurrent networks. Additional exploitation of knowledge sources can help resolve ambiguities. Also, automatic recognition of further traits and states of the subject expressing sentiment such as age, gender, ethnicity, personality, or emotion, and health state can add important information regarding the sentiment expressed.

Make it Multilingual. It seems obvious that multilingualism is an issue for text-based SA. However, it is as well for acoustic analysis, and can in principle even influence video-based SA due to, e. g., varying lip-movements. In fact, languages are often even mixed in real-world statements.

Make it Multicultural. Cross-cultural SA has been researched comparably little, albeit it is clearly of crucial importance. Just as for multilingualism, a key requirement will be sufficient learning data. Then, models can be switched, and transferred across languages and cultures.

3 Conclusion

Seven major requirements were highlighted on the way to truly robust multimodal sentiment analysis in adverse conditions in today's 'big data' era. Beyond these, a number of further issues need to be addressed to best exploit automated sentiment analysis such as provision of meaningful confidence measures and optimal exploitation of 'internal' confidences.

Acknowledgments

The author was supported by the EU's H2020 Programme via the IAs # 644632 (MixedEmotions) and # 645094 (SEWA).

References

[Cambria *et al.*, 2015] E. Cambria, B. Schuller, and Y. Xia. New Avenues in Opinion Mining and Sentiment Analysis. In *Proc. IJCAI 2015*, Buenos Aires, 2015.

Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016, page 1, New York City, USA, July 10, 2016.

WISE: Web-based Interactive Speech Emotion Classification

Sefik Emre Eskimez[⋆], **Melissa Sturge-Apple**[†], **Zhiyao Duan**[⋆] and **Wendi Heinzelman**[⋆]
[⋆]Dept. of Electrical and Computer Engineering
[†]Dept. of Clinical and Social Sciences in Psychology
University of Rochester, Rochester, NY

Abstract

The ability to classify emotions from speech is beneficial in a number of domains, including the study of human relationships. However, manual classification of emotions from speech is time consuming. Current technology supports the automatic classification of emotions from speech, but these systems have some limitations. In particular, existing systems are trained with a given data set and cannot adapt to new data nor can they adapt to different users' notions of emotions. In this study, we introduce WISE, a web-based interactive speech emotion classification system. WISE has a web-based interface that allows users to upload speech data and automatically classify the emotions within this speech using pre-trained models. The user can then adjust the emotion label if the system classification of the emotion does not agree with the user's perception, and this updated label is then fed back into the system to retrain the models. In this way, WISE enables the emotion classification models to be adapted over time. We evaluate WISE by simulating the user interactions with the system using the LDC dataset, which has known, ground-truth labels. We evaluate the benefit of the user feedback enabled by WISE in situations where manually classifying emotions in a large dataset is costly, yet trained models alone will not be able to accurately classify the data.

1 Introduction

Accurately estimating emotions of conversational partners plays a vital role in successful human communication. A social-functional approach to human emotion emphasizes the interpersonal function of emotion for the establishment and maintenance of social relationships [Campos *et al.*, 1989], [Ekman, 1992], [Keltner and Kring, 1998]. According to [Campos *et al.*, 1989] "Emotions are not mere feelings, but rather are processes of establishing, maintaining, or disrupting relations between the person and the internal or external environment, when such relations are significant to the individual." Thus, the expression and recognition of emotions allows the facilitation of social bonds through the conveyance of information about one's internal state, disposition, intentions, and needs.

In many situations, audio is the only recorded data for a social interaction, and estimating emotions from speech becomes a critical task for psychological analysis. Today's technology allows for gathering vast amounts of emotional speech data from the web, yet analyzing this content is impractical. This fact prevents many interesting large-scale investigations.

Given the amount of speech data that proliferates, there have been many attempts to create automatic emotion classification systems. However, the performance of these systems is not as high as necessary in many situations. Many potential applications would benefit from automated emotion classification systems, such as call-center monitoring [Petrushin, 1999; Gupta, 2007], service robot interactions [Park *et al.*, 2009; Liu *et al.*, 2013] and driver assistance systems [Jones and Jonsson, 2005; Tawari and Trivedi, 2010]. Indeed, there are many automated systems today that focus on speech [Sethu *et al.*, 2008; Busso *et al.*, 2009; Rachuri *et al.*, 2010; Bitouk *et al.*, 2010; Stuhlsatz *et al.*, 2011; Yang, 2015]. However, emotion classification accuracy of fully automated systems is still not satisfactory in many practical situations.

In this study, we propose WISE, a web-based interactive speech emotion classification system. This system uses a web-based interface that allows users to easily upload a speech file to the server for emotion analysis, without the need for installing any additional software. Once the speech files are uploaded, the system classifies the emotions using a model trained on previously labeled training samples. Each classification is also associated with a confidence value. The user can either accept or correct the classification, to "teach" the system the user's specific concept of emotions. Over time, the system adapts its emotion classification models to the user's concept, and can increase its classification accuracy with respect to the user's concept of emotions.

The key contribution of our work is that we provide an interactive speech-based emotion analysis framework. This framework combines the machine's computational power with human users' high emotion classification accuracy. Compared to purely manual labeling, it is much more efficient. Compared to fully automated systems, it is much more accurate. This opens up possibilities for large-scale speech emotion analysis with high accuracy.

The proposed framework only considers offline labeling

Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016, pages 2-7, New York City, USA, July 10, 2016.

and returns labels in three categories: emotion, arousal and valance with time codes. To evaluate our system, we have simulated the user-interface interactions in several settings, by providing ground truth labels on behalf of the user. One of the scenarios is designed to be a baseline, with which we can compare the remaining scenarios. In another scenario, we test if the system can adapt to the samples whose speaker is unknown to the system. The next scenario tests how the system's classification confidence of a sample effects the system's accuracy. The full system is available for researchers to use. [1]

The rest of the paper is organized as follows. Section 2 contains a review of the related work. Section 3 describes the WISE web user-interface, while Section 4 explains the automated speech-based emotion recognition system used in this work. We evaluate the WISE system in Section 5, and conclude our work in Section 6.

2 Related Work

All-in-one frameworks for automatic emotion classification from speech, such as EmoVoice [Vogt *et al.*, 2008] and OpenEar [Eyben *et al.*, 2009], are standalone software packages with various capabilities, including audio recording, audio file reading, feature extraction, and emotion classification.

EmoVoice allows the user to create a personal speech-based emotion recognizer, and it can track the emotional state of the user in real-time. Each user records their own speech emotion corpus to train the system, and the system can then be used for real-time emotion classification for the same user. The system outputs the x- and y-coordinates of an arousal-valance coordinate system with time codes. It is reported in [Vogt *et al.*, 2008] that EmoVoice has been used in several systems including humanoid robot-human and virtual agent-human interactions. EmoVoice does not consider user feedback once the classifier is trained, whereas in our system, the user can continually train and improve the system.

OpenEar is an emotion classification multi-platform software package that includes libraries for feature extraction written in C++ and pre-trained models as well as scripts to support model building. One of its main modules is named SMILE (Speech and Music Interpretation by Large-Space Extraction), and it can extract more than 500K features in real-time. The other main module allows external classifiers and libraries such as LibSVM [Chang and Lin, 2011] to be integrated and used in classification. OpenEar also supports popular machine learning frameworks' data formats, such as the Hidden Markov Model Toolkit (HTK) [Young *et al.*, 2006], WEKA [Hall *et al.*, 2009], and scikit-learn for Python [Pedregosa *et al.*, 2011], and therefore allows easy transition between frameworks. OpenEar's capability of batch processing, combined with its advantage in transitioning to other learning frameworks, makes it appealing for large databases.

ANNEMO (ANNotating EMOtions) [Ringeval *et al.*, 2013] is a web-based annotation tool that allows labeling arousal, valence and social dimensions in audio-visual data. The states are represented between -1 and 1, where the user changes the values using a slider. The social dimension is represented by categories rather than numerical values, and those are agreement, dominance, engagement, performance and rapport. No automatic classification/labeling modules are included in ANNEMO.

[1] http://www.ece.rochester.edu/projects/wcng

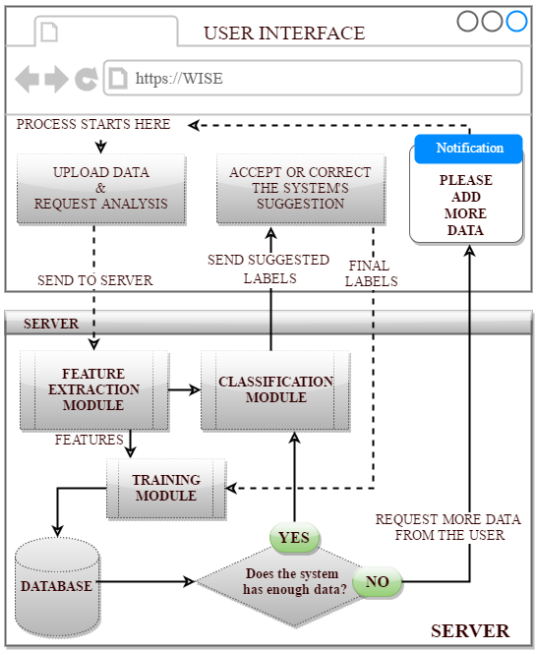


Figure 1: Flow chart showing the operation of WISE.

In contrast, WISE is a web-based system and can be used easily without installing any software, unlike EmoVoice and OpenEar. WISE is similar to ANNEMO in terms of the web-based labeling aspect, however WISE only considers audio data and provides automatic classification as well.

3 Web-based Interaction

Our system's interface, shown in Figure 2, is web-based, allowing easy, secure access and use without installing any other software except a modern browser.

When a user uploads an audio file, the waveform appears on the main screen, allowing the user to select different parts of the waveform. Selected parts can be played and labeled independently. These selected parts will also be added to a list, as shown in the bottom-left side of Figure 2. The user can download this list by clicking on the "save" button in the interface.

The labeling scheme is restricted to three categories: emotion, arousal and valence. Emotion category elements are anger, disgust, fear, happy, neutral, sadness. Arousal category elements are active, passive and neutral, and valance category elements are positive, negative and neutral. Our future work includes adding user defined emotion labels into the system.

The user can request labels from the automated emotion classifier by clicking on the "request label" button. The system then shows suggested labels to the user.

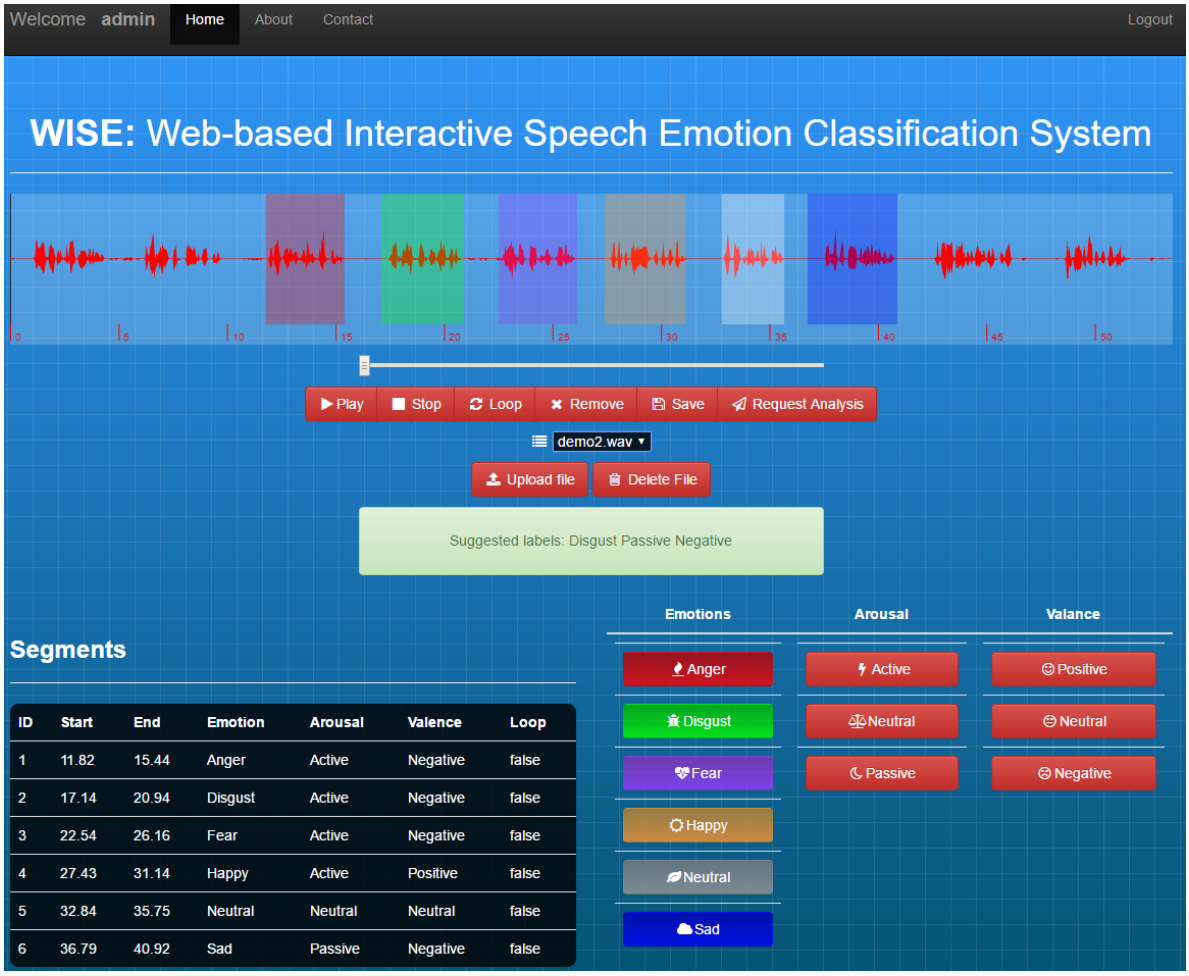


Figure 2: WISE user interface screenshot.

The next section describes the automated speech-based emotion classification system used in WISE.

4 Automated Emotion Classification System

There are various automated speech-based emotion classification systems [Sethu *et al.*, 2008; Busso *et al.*, 2009; Rachuri *et al.*, 2010; Bitouk *et al.*, 2010; Stuhlsatz *et al.*, 2011] that consider different features, feature selection methods, classifiers and decision mechanisms. Our system is based on [Yang, 2015], which provides a confidence value along with the classification label.

4.1 Features

Speech samples are divided into overlapping frames for feature extraction. The window and hop sizes are set to 60 ms and 10 ms, respectively. For every frame that contains speech, the following features are calculated: fundamental frequency (F_0), 12 mel-frequency cepstral coefficients (MFCCs), energy, frequency and bandwidth of first four formants, zero-crossing rate, spectral roll-off, brightness, centroid, spread, skewness, kurtosis, flatness, entropy, roughness, and irregularity, in addition to the derivatives of these features. Statistical values such as minimum, maximum, mean, standard deviation and range (i.e., max-min) are calculated from all frames within the sample. Additionally, speaking rate is calculated over the entire sample. Hence, the final feature vector length is 331.

4.2 Feature Selection

The system employs the support vector machine (SVM) recursive feature elimination method [Guyon *et al.*, 2002]. This approach takes advantage of SVM weights to detect which features are better than others. After the SVM is trained, the features are ranked according to the order of their weights. The last ranked feature is eliminated from the list and the pro-

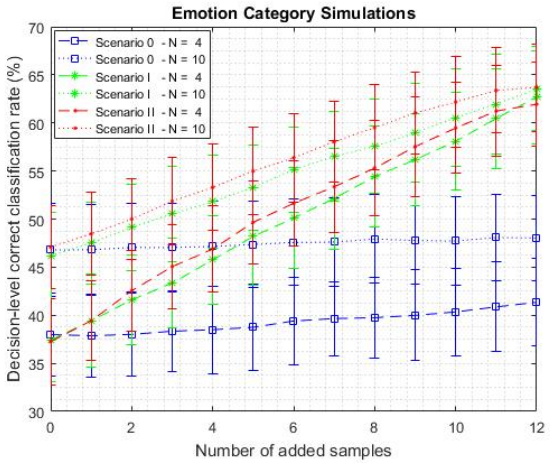


Figure 3: The results of emotion category for Scenarios I-III.

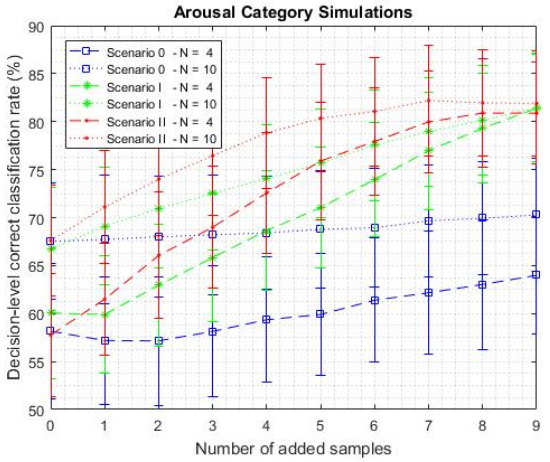


Figure 4: The results of arousal category for Scenarios I-III.

cess starts again, until there are no features left. Features are ranked in reverse order of elimination order. The top 80 best features are chosen to be used in the classification system. Note that in Section 5.2, the features are selected beforehand and not updated when a new sample is added to the system.

4.3 Classifier

Our system uses a one-against-all (OAA) binary SVM with radial basis function (RBF) for each emotion, arousal and valance category element, for a total of 12 SVMs. The trained SVMs calculate confidence scores for any sample that is being classified. The system labels the sample with the class of the binary classifier with maximum classification confidence on the considered sample.

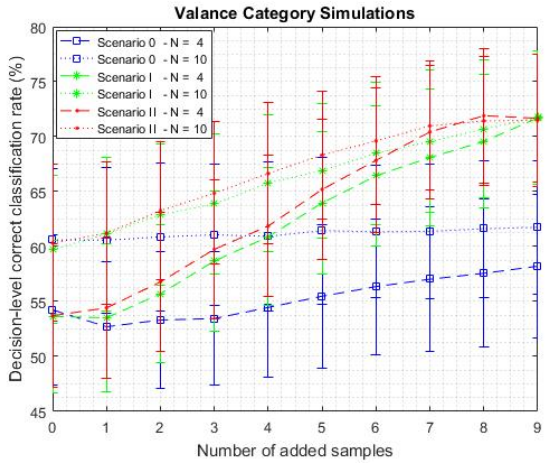


Figure 5: The results of valence category for Scenarios I-III.

5 Evaluation

To evaluate WISE and the benefit of user-assisted labeling of the data, we have simulated user-interface interactions using the LDC database as the source of data for training, validation and testing.

5.1 Dataset

We use the Linguistic Data Consortium (LDC) Emotional Prosody Speech and Transcripts [Liberman *et al.*, 2002] database in our simulations. The LDC database contains samples from 15 emotion categories; however, in our evaluation, we only use 6 of the emotions as listed in Section 3. The LDC database contains acted speech, voiced by 7 professionals, 4 female and 3 male. The transcripts are in English and contain semantically neutral utterances, such as dates and times.

5.2 Simulations

We have simulated user-interface interactions in different scenarios for which WISE can be used to enable user feedback to improve classification accuracy. In these simulations, there are three data groups: training, test and validation. We assume that validation data represents the samples where the user provides the "correct" label. In each iteration, the system evaluates the test data using the current models, and at the end of each iteration, a sample from the validation data is added to the training data to update the models. Next, we describe the different scenarios in detail.

Scenario 0 - Baseline

In this scenario, the data from 1 of the 7 speakers is used for testing, while the remaining 6 speakers' data are used for training and validation. Since only a limited amount of data is available from each speaker in the next scenarios, we also limit the amount of the validation data in this scenario. In this way, the baseline becomes more comparable to the other scenarios.

The training data starts with N samples from each class for each category. For the emotion classification, there are only 2 samples available in each class (emotion) for the validation data. However, the arousal and valance categories have half the number of classes that the emotion category has, therefore, there are 3 samples available in each class that can be used in validation data for these categories. After the data are chosen randomly, the system simulates the interaction process. This process is repeated for all speakers, and the results are averaged over all 7 speakers and 200 trials.

Scenario I

This scenario has the same settings as Scenario 0, except this time, the testing data, as well as the validation data are chosen from a speaker, and the training data is chosen among the remaining 6 speakers' data.

Scenario II

This scenario has the same settings as Scenario I with a single difference: in each round, the validation data has been ordered in ascending order of the classifier's confidence level in classifying them. Therefore in each iteration, the sample, on which the system has the least confidence, is added to the training data from the validation data.

Discussion

Figures 3-5 show the classification accuracy versus the number of added samples for each scenario for the emotion, arousal and valence, respectively. Note that the error bars represent the standard deviation of the results over the 7 speakers and 200 trials.

Scenario I shows the ability of WISE to enable adaptation of the models. In many situations, trained models of automatic systems have no information on the speaker to be classified. The comparison of classification accuracy between Scenario 0 and Scenario I shows that adaptation to unknown data is vital for accurate emotion estimation, as the accuracy increases greatly when data from the new user are added.

For example, in Scenarios I and II, when N is 4 for the emotion category, the system's initial accuracy starts around 37% and increases to approximately 63%, as can be seen in Figure 3, where on the other hand in Scenario 0, accuracy can only increase to approximately 41%. In Scenarios I and II , when N is 10, the classification accuracy starts higher then the previous case, yet with the same number of added samples, they converge to the same percentage. This enables the possibility of using pre-trained models in our system that are trained on available databases.

The results of Scenario II suggest that adding the samples with low classification confidence are slightly more beneficial than adding a sample for which the system already has more confidence. Figures 3-5 show that the classifier in Scenario II converges to a slightly higher classification accuracy than the one in Scenario I. This can be seen especially in the arousal category results.

6 Conclusion

This study introduced and evaluated the WISE system, which is an interactive web-based emotion analysis framework to assist in the classification of human emotion from voice data. The full system is available for the community to use. The evaluation results show that the system can adapt to the user's choices and can increase the future classification accuracy when the speaker of the sample is unknown. Hence, WISE will enable adaptive, large scale emotion classification.

References

[Bitouk *et al.*, 2010] Dmitri Bitouk, Ragini Verma, and Ani Nenkova. Class-level spectral features for emotion recognition. *Speech Commun.*, 52(7):613–625, 2010.

[Busso *et al.*, 2009] C. Busso, S. Lee, and S. Narayanan. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):582–596, May 2009.

[Campos *et al.*, 1989] Joseph J Campos, Rosemary G Campos, and Karen C Barrett. Emergent themes in the study of emotional development and emotion regulation. *Dev Psychol.*, 25(3):394, 1989.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[Ekman, 1992] Paul Ekman. An argument for basic emotions. *Cognition Emotion*, 6(3-4):169–200, 1992.

[Eyben *et al.*, 2009] Florian Eyben, Martin Wllmer, and Bjrn Schuller. openear - introducing the munich open-source emotion and affect recognition toolkit. In *In ACII*, pages 576–581, 2009.

[Gupta, 2007] Purnima Gupta. Two-Stream Emotion Recognition For Call Center Monitoring. In *Interspeech 2007*, 2007.

[Guyon *et al.*, 2002] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46(1-3):389–422, March 2002.

[Hall *et al.*, 2009] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.

[Jones and Jonsson, 2005] Christian Martyn Jones and Ing-Marie Jonsson. Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses. In *Proceedings of the 17th Australia Conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future*, OZCHI '05, pages 1–10, Narrabundah, Australia, Australia, 2005.

[Keltner and Kring, 1998] Dacher Keltner and Ann M Kring. Emotion, social function, and psychopathology. *Rev. Gen. Psychol.*, 2(3):320, 1998.

[Liberman *et al.*, 2002] Mark Liberman, Kelly Davis, M Grossman, N Martey, and J Bell. Emotional prosody speech and transcripts. In *Proc. LDC*, 2002.

[Liu *et al.*, 2013] Chih-Yin Liu, Tzu-Hsin Hung, Kai-Chung Cheng, and Tzuu-Hseng S Li. Hmm and bpnn based speech recognition system for home service robot. In *Advanced Robotics and Intelligent Systems (ARIS), 2013 International Conference on*, pages 38–43. IEEE, 2013.

[Park *et al.*, 2009] J. S. Park, J. H. Kim, and Y. H. Oh. Feature vector classification based speech emotion recognition for service robots. *IEEE Transactions on Consumer Electronics*, 55(3):1590–1596, August 2009.

[Pedregosa *et al.*, 2011] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[Petrushin, 1999] Valery A. Petrushin. Emotion in speech: Recognition and application to call centers. In *In Engr*, pages 7–10, 1999.

[Rachuri *et al.*, 2010] Kiran K Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J Rentfrow, Chris Longworth, and Andrius Aucinas. Emotionsense: a mobile phones based adaptive platform for experimental social psychology research. In *Proc. 12th ACM Int. Conf. on Ubiquitous Computing*, pages 281–290, 2010.

[Ringeval *et al.*, 2013] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, April 2013.

[Sethu *et al.*, 2008] Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Julien Epps. Empirical mode decomposition based weighted frequency feature for speech-based emotion classification. In *Proc. IEEE ICASSP*, pages 5017–5020, 2008.

[Stuhlsatz *et al.*, 2011] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5688–5691, May 2011.

[Tawari and Trivedi, 2010] A. Tawari and M. Trivedi. Speech based emotion classification framework for driver assistance system. In *Intelligent Vehicles Symposium (IV), 2010 IEEE*, pages 174–178, June 2010.

[Vogt *et al.*, 2008] Thurid Vogt, Elisabeth Andr, and Nikolaus Bee. Emovoice - a framework for online recognition of emotions from voice. In *In Proceedings of Workshop on Perception and Interactive Technologies for Speech-Based Systems, Springer, Kloster Irsee*, 2008.

[Yang, 2015] Na Yang. *Algorithms for affective and ubiquitous sensing systems and for protein structure prediction*. PhD thesis, University of Rochester, 2015. http://hdl.handle.net/1802/29666.

[Young *et al.*, 2006] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK, 2006.

Exposing a Set of Fine-Grained Emotion Categories from Tweets

Jasy Liew Suet Yan, Howard R. Turtle
School of Information Studies, Syracuse University
Syracuse, New York, USA
jliewsue@syr.edu, turtle@syr.edu

Abstract

An important starting point in analyzing emotions on Twitter is the identification of a set of suitable emotion classes representative of the range of emotions expressed on Twitter. This paper first presents a set of 48 emotion categories discovered inductively from 5,553 annotated tweets through a small-scale content analysis by trained or expert annotators. We then refine the emotion categories to a set of 28 and test how representative they are on a larger set of 10,000 tweets through crowdsourcing. We describe the two-phase methodology used to expose and refine the set of fine-grained emotion categories from tweets, compare the inter-annotator agreement between annotations generated by expert and novice annotators (crowdsourcing) and show that it is feasible to perform fine-grained emotion classification using gold standard data generated from these two phases. Our main goal is to offer a more representative and finer-grained framework of emotions expressed in microblog text, thus allowing study of emotions that are currently underexplored in sentiment analysis.

1 Introduction

The ways that individuals express themselves in tweets provide windows into their emotional worlds. Twitter, a popular microblogging site with 500 million tweets being sent a day, is particularly rich with emotion expressions. These emotion expressions can be harnessed for sentiment analysis and to build more emotion-sensitive systems. The availability of tweets has paved the way for studies of how emotions expressed on microblogs affect stock market trends [Bollen *et al*., 2011a], relate to fluctuations in social and economic indicators [Bollen *et al*., 2011b], serve as a measure for the population's level of happiness [Dodds and Danforth, 2010], provide situational awareness for both the authorities and the public in the event of disasters [Vo and Collier, 2013], and reflect clinical depression [Park *et al*., 2012].

An important starting point in analyzing emotions on Twitter is the identification of a set of suitable emotion classes. This set of emotion classes should be representative of the emotions expressed in tweets. No consensus has emerged as to how many classes are needed to represent the emotions expressed in text [Farzindar and Inkpen, 2015]. Previous studies have focused on adapting conventional emotion theories from psychology to represent emotions expressed on Twitter and has not attempted to discover the actual range of emotions expressed or how these emotions are actually characterized in tweets. The most commonly used emotion categories are adopted from the basic emotion framework, Ekman's six basic emotions (happiness, sadness, fear, anger, disgust, and surprise) [Ekman, 1971] or Plutchik's eight basic emotions comprising Ekman's six basic emotion, plus the addition of trust and anticipation [Plutchik, 1962].

Instead of borrowing a set of emotion categories from existing emotion theories in psychology, this paper aims to expose a set of categories that are representative of the emotions expressed on Twitter by analyzing the range of emotions humans can reliably detect in microblog text. Our main goal is to offer a more representative and finer-grained framework of emotions expressed in microblog text, thus allowing study of emotions that are currently underexplored in sentiment analysis. In this paper, we address the general research question of what emotions can humans detect in microblog text. We first uncover the set of emotion categories inductively from data and then further refine that set into a manageable set that both humans and machine learning systems are able to reliably detect.

2 Theoretical Background

Generally, we define emotion in text as "a subset of particularly visible and identifiable feelings" [Besnier, 1990; Kagan, 1978] that are expressed in written form through descriptions of expressive reactions (furrowed brow, smile), physiological reactions (increase in heart rate, teeth grinding), cognitions (thoughts of abandonment), behaviors (escape, attack, avoidance) as well as other socially prescribed set of responses [Averill, 1980; Cornelius, 1996]. The classification of emotion in text is largely based on two common models of emotion: 1) the dimensional model, and 2) the categorical model [Calvo and Mac Kim, 2012; Zachar and Ellis, 2012].

Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016, pages 8-14, New York City, USA, July 10, 2016.

The dimensional model organizes emotions into more general dimensions representing the underlying fundamental structure. Emotions can be identified through the composition of two or more independent dimensions [Zachar and Ellis, 2012]. Attempts to identify the dimensions have been conducted through multidimensional scaling of human similarity judgments of emotion expressions based on facial expressions [Abelson and Sermat, 1962], vocal expressions [Green and Cliff, 1975] and emotion terms [Russell, 1978]. The two common dimensions that emerged from these studies are pleasure-displeasure (valence) and degree of arousal (intensity). Similar findings are found in semantic differential studies on emotion terms with the addition of another dimension, dominance-submissiveness [Russell and Mehrabian, 1977]. Valence (also referred to as polarity) classifies emotion as either being positive, negative or neutral [Alm *et al*., 2005; Strapparava and Mihalcea, 2007]. Intensity is somewhat similar to the degree of arousal although it is generally used to measure the strength of the emotion (i.e., very weak to very strong) [Aman and Szpakowicz, 2007]. It can be operationalized as a nominal variable with labels representing varying intensities or measured on a numeric scale.

The categorical model organizes emotions into categories that are formed around prototypes. Each emotion category has a set of distinguishable properties and is assigned a label that best describes the category (e.g., happy, sad and angry). The basic emotion framework follows the categorical model, where emotion is organized and represented using a category system. Each category represents a prototypical emotion. Using a hierarchical classification approach, [Shaver *et al*., 2001] expanded the basic emotions into 25 finer categories through similarity sorting of 135 emotion words. These finer categories are more representative of the emotions that can be expressed using English words.

The dimensional model offers a more coarse-grained representation of emotion while the categorical model can be used to represent emotion at a finer-grained level. In addition, the categorical model uses emotion labels that are more intuitive, thus making recognition of the emotion easier for humans. Therefore, we adopted the categorical model in line with our goal to develop a fine-grained emotion taxonomy for microblog text.

2 Methodology

We used content analysis to identify a stable set of emotion categories that is representative of the range of emotions expressed in tweets. The small-scale content analysis was first conducted (Phase 1) by training a group of annotators to annotate a sample of 5,553 tweets. Three tasks were completed to uncover this set of emotion categories: 1) inductive coding, 2) card sorting, and 3) emotion word rating. In Phase 2, we tested the representativeness of the emotion categories derived from Phase 1 using large-scale content analysis. Annotations were collected through crowdsourcing using Amazon Mechanical Turk (AMT).

2.1 Data Collection

Data consisted of tweets (i.e., microblog posts) retrieved from Twitter. Four different sampling strategies were used to retrieve the tweets to be included in the corpus: random sampling (RANDOM), sampling by topic (TOPIC), and two variations of sampling by user type (SEN-USER and AVG-USER). For the RANDOM sample, nine stopwords (the, be, to, of, and, a, in, that, have) reported to be words most frequently used on Twitter were used to retrieve tweets. Topic sampling was done by retrieving tweets that contain selected topical hashtags or keywords. Sampling by user type retrieved tweets using selected user names (@usernames). One user sample contained tweets retrieved from US Senators (SEN-USER). Tweets from the second user sample were retrieved using randomly selected user names (AVG-USER). Tweets were either retrieved using the Twitter API or acquired from two publicly available data sets: 1) the SemEval 2014 tweet data set [Nakov *et al*., 2013; Rosenthal *et al*., 2014], and 2) the 2012 US presidential elections data set [Mohammad *et al*., 2014]. The data set containing 15,553 tweets received roughly equal contribution from each of the four sampling strategies.

2.2 Phase 1: Small-scale Content Analysis

2.2.1 Task 1: Inductive Coding

We adapted grounded theory [Glaser and Strauss, 1967] to expose a set of fine-grained emotion categories from tweets. This method used inductive coding to derive the classification scheme through observation of content [Potter and Levine-Donnerstein, 1999]. Annotators engaged in three coding activities central to this method: open coding, axial coding, and selective coding [Corbin and Strauss, 2008]. In open coding, annotators read the content of each tweet to capture all possible meanings, and took a first pass at assigning concepts to describe the interpretation of the data. No restriction was posed on analysis in this phase, and minimal instructions were provided to avoid predisposing annotators. Axial coding then involved the process of drawing the relationships between concepts and categories. Based on their knowledge of emotion, annotators started with a set of self-defined emotion tags. They then met in groups with the primary researcher to start drawing relationships between different emotion tags suggested by individuals in the group. Emotion tags were examined, accepted, modified, and discarded. Discrete emotion categories started to form in this phase, and were systematically applied to more data. Annotators switched back and forth between axial coding and open coding until a stable set of categories was identified. Finally, selective coding represented an integration phase where the identified discrete categories were further developed, defined and refined under a unifying theme of emotion. Annotators then continued to validate the classification scheme by applying and refining it on more data until no new category emerged.

Graduate students who were interested in undertaking the task as part of a class project (e.g., Natural Language Processing course) or to gain research experience in content

analysis (e.g., independent study) were recruited as annotators. Annotators were not expected to possess special skills except for the required abilities to read and interpret English text. A total of eighteen annotators worked on the annotation task over a period of ten months. To derive an emotion framework based on collective knowledge, each tweet was annotated by at least three annotators. Thus, annotators were divided into groups of at least three. Each group was assigned to work on one of the four samples.

All the annotators went through the same training procedures to reduce as much as possible the variation among different individuals. Each annotator first attended a one hour training session to discuss the concept of emotion with the researcher and to receive instructions on how to perform annotations of the tweets. Annotators were not given any emotion categories and were asked to suggest the best-fitting emotion tags or labels to describe the emotion expressed in each tweet (Example 1). For tweets containing multiple emotions, annotators were asked to first identify the primary emotion expressed in the tweet, and then also include the other emotions observed (Example 2).

Example 1: Alaska is so proud of our Spartans! The 4-25 executed every mission in Afghanistan with honor & now, they're home http://t.co/r8pLpnud **[Pride]**

Example 2: Saw Argo yesterday, a movie about the 1979 Iranian Revolution. Chilling, sobering, and inspirational at the same time. **[Inspiration, Fear]**

Annotation was done in an iterative fashion. In the first iteration, also referred to as the training round, all annotators annotated the same sample of 300 tweets from SEN-USER. Upon completing the training round, annotators were assigned to annotate at least 1,000 tweets from one of the four samples (RANDOM, TOPIC, AVG-USER or SEN-USER) in subsequent iterations. Every week, annotators worked independently on annotating a subset of 150 – 200 tweets but met with the researcher in groups to discuss disagreements, and 100% agreement for emotion tag was achieved after discussion. In these weekly meetings, the researcher also facilitated the discussions among annotators working on the same sample to merge, remove, and refine suggested emotion tags. Output of Task 1 included 4,010 annotated tweets in the gold standard corpus and 246 emotion tags.

2.2.2 Task 2: Card Sorting

Some of the 246 emotion tags were simply morphological variations and many were semantically similar. Task 2 served as an intermediate step to refine the emotion tags emerging from data into a more manageable set of higher level emotion categories. Annotators were asked to perform a card sorting exercise in different teams to group emotion tags that are variants of the same root word or semantically similar into the same category. Annotators were divided into 5 teams, and each team received a pack of 1' x 5' cards containing only the emotion tags used by the all members in their respective teams.

Each team consisted of 2 - 3 members who worked on the same sample. Teams were instructed to follow the four-step procedures described below:

a) Group all the emotion tags into categories. Members were allowed to create a "Not Emotion" category if needed.
b) Create a name for the emotion category. Collectively pick the most descriptive emotion tag or suggest a new name to represent each category.
c) Group all the emotion categories based on valence: positive, negative and neutral.
d) Match emotion categories generated from other team's card sorting activity to the emotion categories proposed by your team.

Team	Sample	Number of Emotion Categories			
		Positive	Negative	Neutral	Total
G1	SEN-USER	8	13	2	23
G2	TOPIC	16	14	5	35
G3	TOPIC	16	18	8	42
G4	AVG-USER	14	18	15	47
G5	RANDOM	14	16	9	39

Table 1: Number of categories proposed by each card sorting team

Members in the same team were allowed to discuss their decisions with each other during the card sorting exercise with minimal intervention from the researcher. The session concluded when all members completed the four-step procedure and reached a consensus on final groupings of the emotion tags. No limit was placed on the number of categories or the number of emotion tags within each category so the number of categories proposed varied across the five teams as shown in Table 1. Some teams decided to put the emotion tags into fewer higher-level categories, while others who chose to capture more subtle emotions generated more emotion categories. Finally, the researcher merged, divided, and verified the final emotion categories to be included in the classification scheme.

Once the final 48 emotion categories shown in Table 2 were identified (see Emotion-Category-48 column), the original emotion tag labels generated from the open coding exercise were systematically replaced by the appropriate emotion category labels. Annotators then incrementally annotated more tweets (150 - 200 tweets per round) to ensure that a point of saturation was reached. No new emotion category emerged from data in this coding phase. Another 1,543 annotated tweets with gold labels were added to the corpus.

2.2.3 Task 3: Emotion Word Rating

We found it methodologically challenging and time consuming to provide rigorous training to a large number of annotators in order to grow the size of the corpus with 48 emotion categories. A word rating study was conducted as a systematic method to merge and distill the number of categories into a more manageable set. The motivation behind the word rating study came from prior studies showing that emotion words with greater similarity tend to be in close proximity to one another on a two-dimensional

pleasure and degree of arousal space [Russell, 1980]. In order to plot our emotion categories in this two-dimensional space, we collected the pleasure and arousal ratings for each emotion category. A set of 50 emotion words were selected for the emotion rating task. We included the 48 emotion category names and added 2 emotion words that were deemed to be more appropriate category names than the ones determined by the annotators in Task 2. These two emotion words were *"longing"* for the category *"yearning"* and *"torn"* for *"ambivalence"*.

To obtain a complete set of pleasure and arousal ratings for our set of 50 emotion words, we conducted an emotion word rating study on AMT. We adapted the instrument that was used in [Bradley & Lang, 1999] to collect the ratings. We implemented the study using exactly the same 9-point scale for the pleasure and arousal ratings. The validity of the scales are described in [Bradley & Lang, 1994]. The same set of instructions was reused but modified to fit the crowdsourcing context.

Human raters were recruited from the pool of workers available on AMT. The rating instrument was offered to the workers via a Human Intelligence Task (HIT), and workers received payment of US$ 0.20 upon completion and approval of the HIT. HITs were restricted to workers in the US to increase the likelihood that ratings came from native English speakers. Each respondent first read the instructions on how to use the pleasure and arousal scales. Respondents were then instructed to make a pleasure rating and an arousal rating for each of the 50 emotion words.

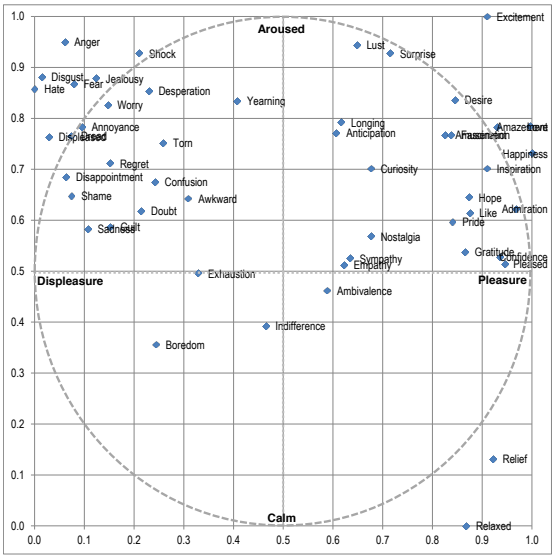


Figure 1: Two-dimensional pleasure and arousal plot for 50 emotion words based on AMT ratings (x-axis represents pleasure, y-axis represents arousal)

After removing incomplete and rejected responses, mean rating and standard deviation were computed from 76 usable responses.

Figure 1 shows the plot for all 50 emotion words based on AMT ratings normalized using feature scaling. Emotion categories that are semantically-related and relatively close in proximity to one another on the plot are merged. The merge process involved some subjective decision and reduced the number of emotion categories from 48 to the final set of 28 shown in Table 2 (see Emotion-Category-28 column). Category name "ambivalence" was substituted by its more descriptive member term, "torn" and "yearning" was substituted by "longing". Also, two emotion categories from the original 48, "desire" and "lust" were dropped altogether from the final set of 28 because it is not clear that they should be considered separate emotional states [Ortony and Turner, 1990]. Based on their conceptualization in our annotation scheme, they were considered to be more general feelings of wanting rather than distinct emotional states. Finally, the 48 emotion category labels in the corpus were systematically replaced by the corresponding 28 emotion category labels.

The set of 28 categories is derived from the corpus and is a "good" representation of the set of emotions expressed therein. It is substantially more refined than the traditional 5 to 8 category set yet is small enough that human annotators are comfortable with the distinctions.

2.3 Phase 2: Large-scale Content Analysis

Manual annotations for an additional 10,000 tweets were obtained using AMT in Phase 2. For emotion tag, workers were given a set of 28 emotion categories to choose from plus an "other" option with a text box so they could suggest a new emotion tag where none of the listed emotion category was applicable. The order in which the emotion categories were presented to the workers was randomized across the four samples in order to control for order effect. If a tweet was flagged as containing multiple emotions, annotators were asked to provide all relevant emotion tags.

Recruitment of workers was done through Human Intelligence Tasks (HITs) on the online AMT platform. AMT workers must fulfill at least the basic requirement of being able to read and understand English text. We set the HIT approval rate for all requesters' HITs to greater than or equal to 95% and the number of HITs approved to greater than or equal to 1000 to increase the probability of recruiting first-rate workers.

In the design of the HIT, workers were provided clear and simple instructions describing the task, the annotation site link, as well as a batch id required to retrieve a subset of 30 tweets to work on. Of the 30 tweets in each HIT, 25 were new tweets and 5 were gold standard tweets intended to be used for quality control. Each HIT was assigned to three different annotators. Each HIT bundled a different subset of 30 tweets so a worker could attempt more than one HIT. Workers were paid US$ 0.50 for every completed and approved HIT containing 30 tweets.

Emotion-Category-28	Emotion-Category-48	Emotion-Category-28	Emotion-Category-48
Admiration	Admiration	Hate	Hate, Disgust
Amusement	Amusement	Hope	Hope
Anger	Anger, Annoyance, Displeased, Disappointment	Indifference	Indifference
Boredom	Boredom	Inspiration	Inspiration
Confidence	Confidence	Jealousy	Jealousy
Curiosity	Curiosity	Longing	*Longing, Nostalgia
Desperation	Desperation	Love	Love, Like
Doubt	Doubt, Confusion, *Torn	Pride	Pride
Excitement	Excitement, Anticipation	Regret	Regret, Guilt
Exhaustion	Exhaustion	Relaxed	Relaxed, Relief
Fascination	Fascination, Amazement	Sadness	Sadness
Fear	Fear, Dread, Worry	Shame	Shame, Awkward
Gratitude	Gratitude	Surprise	Surprise, Shock
Happiness	Happiness, Pleased	Sympathy	Sympathy, Empathy

Table 2: Mapping between the final set of 28 emotion categories to the original set of 48 (category names preceded by * were modified)

Phase	Phase 1				Phase 2				Phase 1 + Phase 2			
Category	**n**	**P**	**R**	**F1**	**n**	**P**	**R**	**F2**	**n**	**P**	**R**	**F1**
Admiration	158	0.417	0.190	0.261	245	0.328	0.155	0.211	403	0.370	0.201	0.260
Amusement	237	0.744	0.515	0.608	423	0.888	0.617	0.728	660	0.869	0.645	0.741
Anger	444	0.288	0.203	0.238	757	0.495	0.346	0.407	1201	0.478	0.321	0.384
Boredom	12	0.400	0.167	0.235	36	0.714	0.417	0.526	48	0.818	0.375	0.514
Confidence	19	0.000	0.000	0.000	91	0.286	0.088	0.134	110	0.303	0.091	0.140
Curiosity	30	0.586	0.567	0.576	63	0.591	0.413	0.486	93	0.638	0.548	0.590
Desperation	8	1.000	0.125	0.222	50	0.417	0.100	0.161	58	0.500	0.069	0.121
Doubt	50	0.125	0.020	0.034	108	0.256	0.102	0.146	158	0.269	0.089	0.133
Excitement	265	0.457	0.377	0.413	421	0.675	0.463	0.549	686	0.655	0.474	0.550
Exhaustion	10	0.000	0.000	0.000	39	0.706	0.308	0.429	49	0.611	0.224	0.328
Fascination	54	0.417	0.185	0.256	150	0.587	0.360	0.446	204	0.553	0.309	0.396
Fear	77	0.240	0.078	0.118	162	0.556	0.216	0.311	239	0.491	0.230	0.313
Gratitude	221	0.943	0.905	0.924	300	0.913	0.877	0.895	521	0.928	0.914	0.921
Happiness	778	0.589	0.500	0.541	1009	0.596	0.477	0.530	1787	0.622	0.506	0.558
Hate	63	0.778	0.444	0.566	129	0.812	0.535	0.645	192	0.788	0.542	0.642
Hope	187	0.660	0.508	0.574	335	0.781	0.564	0.655	522	0.781	0.580	0.666
Indifference	28	0.500	0.071	0.125	40	0.308	0.100	0.151	68	0.235	0.059	0.094
Inspiration	21	0.923	0.571	0.706	54	0.731	0.352	0.475	75	0.816	0.413	0.549
Jealousy	5	1.000	0.400	0.571	29	0.846	0.379	0.524	34	0.765	0.382	0.510
Longing	41	0.545	0.146	0.231	80	0.487	0.238	0.319	121	0.529	0.306	0.387
Love	234	0.608	0.444	0.514	447	0.645	0.538	0.587	681	0.659	0.519	0.581
Pride	85	0.817	0.682	0.744	128	0.907	0.688	0.782	213	0.862	0.676	0.758
Regret	49	0.500	0.102	0.169	104	0.571	0.308	0.400	153	0.514	0.242	0.329
Relaxed	26	0.200	0.038	0.065	51	0.550	0.216	0.310	77	0.737	0.182	0.292
Sadness	158	0.609	0.335	0.433	363	0.612	0.444	0.514	521	0.650	0.461	0.539
Shame	26	0.600	0.231	0.333	64	0.545	0.281	0.371	90	0.622	0.311	0.415
Surprise	93	0.342	0.140	0.198	173	0.627	0.301	0.406	266	0.556	0.278	0.371
Sympathy	35	0.813	0.371	0.510	66	0.625	0.379	0.472	101	0.705	0.426	0.531
None	2637				5047				7684			
Macro-avg		**0.539**	**0.297**	**0.363**		**0.609**	**0.366**	**0.449**		**0.619**	**0.370**	**0.450**
Micro-avg		**0.580**	**0.400**	**0.474**		**0.647**	**0.440**	**0.524**		**0.656**	**0.455**	**0.537**

Table 3: Precision, recall and F1 of SMO classifiers across P1, P2 and P1+P2

3 Inter-annotator Agreement

Table 4 presents the inter-annotator agreement for 28 emotion categories based on tweets with three annotations. Overall α across Phase 1 and Phase 2 is 0.43. Mean α for 28 emotion categories in Phase 1 is 0.50. Emotion annotation especially at a fine-grained level is a subjective and difficult task. It is possible to generate reliable data when annotators are given sufficient training. With limited training, α in Phase 2 decreases almost by half to 0.28.

Sample	EmoCat-28		
	%	κ	α
Phase 1	66	0.50	0.50
Phase 2	51	0.28	0.28
Phase 1+Phase 2	61	0.43	0.43

Table 4: Inter-annotator agreement (percent agreement, Fleiss' κ and Krippendorff's α)

For Phase 1, all disagreements were first resolved through discussion with expert annotators. Essentially, expert annotators achieved 100% agreement in Phase 1. In Phase 2, about one third of the tweets had full agreement for emotion tag among all annotators (32%). To avoid throwing away any data, the researcher manually reviewed all annotations and resolved the disagreements. Such effort was deemed necessary to reduce as much noise as possible in the corpus, and to ensure that the classification schemes were applied consistently across the two phases of data collection. Similar to the Phase 1, each tweet in Phase 2 was assigned final labels for emotion category.

4 Emotion Distribution

Slightly over half (51%) of the tweets contain emotion. Table 3 shows imbalance in the frequencies of the emotion categories. Of the 28 emotion categories, the full corpus (Phase 1 and Phase 2) contains the highest instances of *happiness* (12%) and the lowest instances of jealousy (0.2%). Only 9 categories have less than 100 instances. The frequency distribution of the emotion categories in Phase 1, Phase 2 and Phase 1 + Phase 2 are roughly similar.

The corpus contains a significant portion of tweets tagged with a single emotion category (92%) and only 8% of tweets tagged with more than one emotion category. Although tweets containing multiple emotions represent only 8% of the corpus, including such tweets in the corpus leads to over 40% overall increase in the number of positive examples (i.e., instances of an emotion category).

5 Comparing Machine Learning Results from Phase 1 and Phase 2

Since a tweet might be assigned multiple emotion categories, we frame the problem as a multi-label classification task. A separate binary classifier was built for each emotion category to detect if an emotion category were present or absent in a tweet (emotion X or not emotion X).

We conducted a wide range of classification experiments to better understand the impact of classifier and feature set selection on classification accuracy [Liew, 2016]. We present here results for a single representative selection: Sequential Minimal Optimization (SMO), an SVM variant [Platt, 1998] trained with features that include unigrams occurring three or more times in the corpus that are stemmed and lowercased. Classifiers were evaluated using ten-fold cross validation.

The precision, recall and F1 for SMO across Phase 1, Phase 2 and Phase 1 + Phase 2 are shown in Table 3. A general upward trend in precision (P), recall (R) and F1 are observed across the three data sets. There are two key takeaways from our preliminary experiments. First, using the combined data from P1 and P2 generally yields higher performance than using P1 or P2 data alone. For a majority of the emotion categories, the classifiers used for emotion classification achieved similar performance using gold standard data generated Phase 1 and Phase 2 respectively. Second, classifiers provided with more training examples usually produce higher overall performance as evidenced by higher F1 when larger data sets are used. The results for individual emotion categories shows that more data does not always leads to higher performance. The classifiers may behave differently depending on the linguistic characteristics of the category. More experiments will be conducted in future work to identify the salient linguistic features for each emotion category.

6 Conclusion

We describe a two-phase methodology to uncover a set of 28 emotion categories representative of the emotions expressed in tweets. There are two main contributions: 1) the introduction an emotion taxonomy catered for emotion expressed in text and 2) the development of a gold standard corpus that can be used to train and evaluate more fine-grained emotion classifiers.

The set of 28 emotions is derived using an integrative view of emotion and grounded on linguistic expressions of emotion in text. In Phase 1, inductive coding was first used to expose a set of emotion categories from 5,553 tweets. The categories were then further merged and refined using card sorting and emotion word rating. In Phase 2, we then tested the representativeness of the emotion categories on a larger data set of 10,000 tweets using crowdsourcing. No new emotion categories emerged from Phase 2, indicating that the 28 emotion categories are sufficient to capture the richness of emotional experiences expressed in tweets. However, the classifiers perform poorly on some categories such as *confidence*, *desperation*, *doubt* and *indifference*. We intend to perform a closer examination of the low performing categories to determine if they should be removed.

Acknowledgments

We thank the annotators who volunteered in performing the annotation task. We are grateful to Dr. Elizabeth D. Liddy for her insights in the study.

References

[Abelson and Sermat, 1962] Robert Abelson, and Vello Sermat. Multidimensional Scaling of Facial Expressions. *Journal of Experimental Psychology*, 63(6):546–554, 1962.

[Alm et al., 2005] Cecilia Alm, Dan Roth, and Richard Sproat. Emotions from Text: Machine Learning for Text-Based Emotion Prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 579–586, Stroudsburg, PA, USA, 2005.

[Aman and Szpakowicz, 2007] Saima Aman, and Stan Szpakowicz. Identifying Expressions of Emotion in Text. In *Text, Speech and Dialogue*, pages 196–205, 2007.

[Averill, 1980] James R. Averill. A Constructivist View of Emotion. *Emotion: Theory, Research, and Experience*, 1:305–339, Academic Press, New York, 1980.

[Besnier, 1990] Niko Besnier. Language and Affect. *Annual Review of Anthropology*, 19:419–451, 1990.

[Bollen *et al*., 2011a] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter Mood Predicts the Stock Market. *Journal of Computational Science*, 2(1):1–8, 2011.

[Bollen *et al*., 2011b] Johan Bollen, Alberto Pepe, and Huina Mao. Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, pages 450–53, 2011.

[Bradley *et al*., 1994] Margaret M. Bradley, and Peter J. Lang. Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1): 49–59, 1994.

[Bradley and Lang, 1999] Margaret M. Bradley, and Peter J. Lang. Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. University of Florida: Technical Report C-1, The Center for Research in Psychophysiology, 1999.

[Calvo and Mac Kim, 2012] Rafael A. Calvo, and Sunghwan Mac Kim. Emotions in Text: Dimensional and Categorical Models. *Computational Intelligence*, 29(3):527–43, 2012.

[Corbin and Strauss, 2008] Juliet Corbin, and Anselm Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage, 2008.

[Cornelius, 1996] Randolph R. Cornelius. *The Science of Emotion: Research and Tradition in the Psychology of Emotions*. Upper Saddle River, Prentice Hall, New Jersey, 1996.

[Dodds and Danforth, 2010] Peter S. Dodds, and Christopher M. Danforth. Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents. *Journal of Happiness Studies*, 11(4):441–56, 2010.

[Ekman, 1971] Paul Ekman. Universals and Cultural Differences in Facial Expressions of Emotion. *Nebraska Symposium on Motivation*, 19:207–83, 1971.

[Farzindar and Inkpen, 2015] Atefeh Farzindar, and Diana Inkpen. Natural Language Processing for Social Media. *Synthesis Lectures on Human Language Technologies*, 8(2):1–166, 2015.

[Glaser and Strauss, 1967] Barney G. Glaser, and Anselm L. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*, Aldine Publishing, Chicago 1967.

[Green and Cliff, 1975] Rex S. Green, and Norman Cliff. Multidimensional Comparisons of Structures of Vocally and Facially Expressed Emotion. *Perception & Psychophysics*, 17(5):429–438, 1975.

[Kagan, 1978] Jerome Kagan. On Emotion and Its Development: A Working Paper. In *The Development of Affect*, pages 11–41, Genesis of Behavior 1, 1978.

[Liew, 2016] Jasy Liew Suet Yan. *Fine-Grained Emotion Detection in Microblog Text*. Syracuse, NY, USA: Syracuse University, 2016.

[Mohammad *et al*., 2014] Saif Mohammad, Xiaodan Zhu, and Joel Martin. Semantic Role Labeling of Emotions in Tweets." In *Proceedings of the ACL 2014 Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media*, pages 32–41, Baltimore, MD, USA, 2014.

[Nakov *et al*., 2013] Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. SemEval-2013 Task 2: Sentiment Analysis in Twitter." In *Proceedings of the 7th International Workshop on Semantic Evaluation* , 2:312–320, 2013.

[Ortony and Turner, 1990] Andrew Ortony, and Terence J. Turner. What's Basic about Basic Emotions? *Psychological Review*, 97(3):315–331, 1990.

[Park *et al*., 2012] Minsu Park, Chiyoung Cha, and Meeyoung Cha. Depressive Moods of Users Portrayed in Twitter. In *Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics*, pages 1–8, 2012.

[Platt, 1998] John C. Platt. Fast Training of Support Vector Machines Using Sequential Minimal Optimization. In *Advances in Kernel Methods*, pages 41–65, MIT Press, 1998.

[Plutchik, 1962] Robert Plutchik. *The Emotions: Facts, Theories, and a New Model*. Studies in Psychology, Random House, New York, 1962.

[Potter and Levine-Donnerstein, 1999] W. James Potter, and Deborah Levine-Donnerstein. Rethinking Validity and Reliability in Content Analysis. *Journal of Applied Communication Research*, 27(3):258–284, 1999.

[Rosenthal *et al*., 2014] Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. Semeval-2014 Task 9: Sentiment Analysis in Twitter." In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 73–80. Dublin, Ireland, 2014.

[Russell, 1978] James A. Russell. Evidence of Convergent Validity on the Dimensions of Affect. *Journal of Personality and Social Psychology*, 36(10):1152–1168, 1978.

[Russell, 1980] James A. Russell. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[Russell and Mehrabian, 1977] James A. Russell, and A. Mehrabian. Evidence for a Three-Factor Theory of Emotions. *Journal of Research in Personality*, 11(3):273–294, 1977.

[Shaver et al., 2001] Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'Connor. Emotion Knowledge: Further Exploration of a Prototype Approach. In *Emotions in Social Psychology*, pages 26–56. Psychology Press, 2001.

[Strapparava and Mihalcea, 2007] Carlo Strapparava, and Rada Mihalcea. Semeval-2007 Task 14: Affective Text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Prague, 2007.

[Vo and Collier, 2013] Bao-Khanh H. Vo, and Nigel Collier. Twitter Emotion Analysis in Earthquake Situations. *International Journal of Computational Linguistics and Applications*, 4(1):159–173, 2013.

[Zachar and Ellis, 2012] Peter Zachar, and Ralph D. Ellis. *Categorical versus Dimensional Models of Affect: A Seminar on the Theories of Panksepp and Russell*. Vol. 7. John Benjamins Publishing Company, 2012.

SMILE: Twitter Emotion Classification using Domain Adaptation

Bo Wang **Maria Liakata** **Arkaitz Zubiaga** **Rob Procter** **Eric Jensen**
Department of Computer Science
University of Warwick
Coventry, UK
{bo.wang, m.liakata, e.jensen}@warwick.ac.uk

Abstract

Despite the widely spread research interest in social media sentiment analysis, sentiment and emotion classification across different domains and on Twitter data remains a challenging task. Here we set out to find an effective approach for tackling a cross-domain emotion classification task on a set of Twitter data involving social media discourse around arts and cultural experiences, in the context of museums. While most existing work in domain adaptation has focused on feature-based or/and instance-based adaptation methods, in this work we study a model-based adaptive SVM approach as we believe its flexibility and efficiency is more suitable for the task at hand. We conduct a series of experiments and compare our system with a set of baseline methods. Our results not only show a superior performance in terms of accuracy and computational efficiency compared to the baselines, but also shed light on how different ratios of labelled target-domain data used for adaptation can affect classification performance.

1 Introduction

With the advent and growth of social media as a ubiquitous platform, people increasingly discuss and express opinions and emotions towards all kinds of topics and targets. One of the topics that has been relatively unexplored in the scientific community is that of emotions expressed towards arts and cultural experiences. A survey conducted in 2012 by the British TATE Art Galleries found that 26 percent of the respondents had posted some kind of content online, such as blog posts, tweets or photos, about their experience in the art galleries during or after their visit [Villaespesa, 2013]. When cultural tourists share information about their experience in social media, this real-time communication and spontaneous engagement with art and culture not only broadens its target audience but also provides a new space where valuable insight shared by its customers can be garnered. As a result museums, galleries and other cultural venues have embraced social media such as Twitter, and actively used it to promote their exhibitions, organise participatory projects and/or create initiatives to engage with visitors, collecting valuable opinions and feedback (e.g. museum tweetups). This gold mine of user opinions has sparked an increasing research interest in the interdisciplinary field of social media and museum study [Fletcher and Lee, 2012; Villaespesa, 2013; Drotner and Schrøder, 2014].

We have also seen a surge of research in sentiment analysis with over 7,000 articles written on the topic [Feldman, 2013], for applications ranging from analyses of movie reviews [Pang and Lee, 2008] and stock market trends [Bollen *et al.*, 2011] to forecasting election results [Tumasjan *et al.*, 2010]. Supervised learning algorithms that require labelled training data have been successfully used for in-domain sentiment classification. However, cross-domain sentiment analysis has been explored to a much lesser extent. For instance, the phrase "light-weight" carries positive sentiment when describing a laptop but quite the opposite when it is used to refer to politicians. In such cases, a classifier trained on one domain may not work well on other domains. A widely adopted solution to this problem is domain adaptation, which allows building models from a fixed set of source domains and deploy them into a different target domain. Recent developments in sentiment analysis using domain adaptation are mostly based on feature-representation adaptation [Blitzer *et al.*, 2007; Pan *et al.*, 2010; Bollegala *et al.*, 2011], instance-weight adaptation [Jiang and Zhai, 2007; Xia *et al.*, 2014; Tsakalidis *et al.*, 2014] or combinations of both [Xia *et al.*, 2013; Liu *et al.*, 2013]. Despite its recent increase in popularity, the use of domain adaptation for sentiment and emotion classification across topics on Twitter is still largely unexplored [Liu *et al.*, 2013; Tsakalidis *et al.*, 2014; Townsend *et al.*, 2014].

In this work we set out to find an effective approach for tackling the cross-domain emotion classification task on Twitter, while also furthering research in the interdisciplinary study of social media discourse around arts and cultural experiences[1]. We investigate a model-based adaptive-SVM approach that was previously used for video concept detection [Yang *et al.*, 2007] and compare with a set of domain-dependent and domain-independent strategies. Such a model-based approach allows us to directly adapt existing models to the new target-domain data without having to generate domain-dependent features or adjusting weights for each of

[1]SMILE project: `http://www.culturesmile.org/`

Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016, pages 15-21, New York City, USA, July 10, 2016.

the training instances.We conduct a series of experiments and evaluate the proposed system[2] on a set of Twitter data about museums, annotated by three annotators from the social sciences. The aim is to maximise the use of the base classifiers that were trained from a general-domain corpus, and through domain adaptation minimise the classification error rate across 5 emotion categories: *anger*, *disgust*, *happiness*, *surprise* and *sadness*. Our results show that adapted SVM classifiers achieve significantly better performance than out-of-domain classifiers and also suggest a competitive performance compared to in-domain classifiers. To the best of our knowledge this is the first attempt at cross-domain emotion classification for Twitter data.

2 Related Work

Most existing approaches can be classified into two categories: feature-based adaptation and instance-based adaptation. The former seek to construct new adaptive feature representations that reduce the difference between domains, while the latter aims to sample and re-weight source domain training data for use in classification within the target domain.

With respect to feature domain adaptation, [Blitzer *et al.*, 2007] applied structural correspondence learning (SCL) algorithm for cross-domain sentiment classification. SCL chooses a set of *pivot features* with highest mutual information to the domain labels, and uses these pivot features to align other features by training N linear predictors. Finally it computes singular value decomposition (SVD) to construct low-dimensional features to improve its classification performance. A small amount of target domain labelled data is used to learn to deal with misaligned features from SCL. [Townsend *et al.*, 2014] found that SCL did not work well for cross-domain adaptation of sentiment on Twitter due to the lack of mutual information across the Twitter domains and uses subjective proportions as a backoff adaptation approach. [Pan *et al.*, 2010] proposed to construct a bipartite graph from a co-occurrence matrix between domain-independent and domain specific features to reduce the gap between different domains and use spectral clustering for feature alignment. The resulting clusters are used to represent data examples and train sentiment classifiers. They used mutual information between features and domains to classify domain-independent and domain specific features, but in practice this also introduces mis-classification errors. [Bollegala *et al.*, 2011] describes a cross-domain sentiment classification approach using an automatically created sentiment sensitive thesaurus. Such a thesaurus is constructed by computing the point-wise mutual information between a lexical element u and a feature as well as relatedness between two lexical elements. The problem with these feature adaptation approaches is that they try to connect domain-dependent features to known or common features under the assumption that parallel sentiment words exist in different domains, which is not necessarily applicable to various topics in tweets [Liu *et al.*, 2013]. [Glorot *et al.*, 2011] proposes a deep learning system to extract features that are highly beneficial for the domain adaptation of sentiment classifiers, under the intuition that deep learning algorithms learn intermediate concepts (between raw input and target) and these intermediate concepts could yield better transfer across domains.

When it comes to instance adaptation, [Jiang and Zhai, 2007] proposes an instance weighting framework that prunes "misleading" instances and approximates the distribution of instances in the target domain. Their experiments show that by adding some labelled target domain instances and assigning higher weights to them performs better than either removing "misleading" source domain instances using a small number of labelled target domain data or bootstrapping unlabelled target instances. [Xia *et al.*, 2014] adapts the source domain training data to the target domain based on a logistic approximation. [Tsakalidis *et al.*, 2014] learns different classifiers on different sets of features and combines them in an ensemble model. Such an ensemble model is then applied to part of the target domain test data to create new training data (i.e. documents for which different classifiers had the same predictions). We include this ensemble method as one of our baseline approaches for evaluation and comparison.

In contrast with most cross-domain sentiment classification works, we use a model-based approach proposed in [Yang *et al.*, 2007], which directly adapts existing classifiers trained on general-domain corpora. We believe this is more efficient and flexible [Yang and Hauptmann, 2008] for our task. We evaluate on a set of manually annotated tweets about cultural experiences in museums and conduct a finer-grained classification of emotions conveyed (i.e. *anger*, *disgust*, *happiness*, *surprise* and *sadness*).

3 Datasets

We use two datasets, a source-domain dataset and a target-domain dataset, which enables us to experiment on domain adaptation. The source-domain dataset we adopted is the general-domain Twitter corpus created by [Purver and Battersby, 2012], which was generated through distant supervision using hashtags and emoticons associated with 6 emotions: anger, disgust, fear, happiness, surprise and sadness.

Our target-domain dataset that allows us to perform experiments on emotions associated with cultural experiences consists of a set of tweets pertaining to museums. A collection of tweets mentioning one of the following Twitter handles associated with British museums was gathered between May 2013 and June 2015: @camunivmuseums, @fitzmuseum_uk, @kettlesyard, @maacambridge, @iciabath, @thelmahulbert, @rammuseum, @plymouthmuseum, @tateliverpool, @tate_stives, @nationalgallery, @britishmuseum, @_thewhitechapel. These are all museums associated with the SMILES project. A subset of 3,759 tweets was sampled from this collection for manual annotation. We developed a tool for manual annotation of the emotion expressed in each of these tweets. The options for the annotation of each tweet included 6 different emotions; the six Ekman emotions as in [Purver and Battersby, 2012], with the exception of 'fear' as it never featured in the context of tweets about museums. Two extra annotation options were included to indicate that a tweet should have *no code*, indicating that a tweet was

[2]The code can be found at `http://bit.ly/1WHup4b`

not conveying any emotions, and *not relevant* when it did not refer to any aspects related to the museum in question. The annotator could choose more than one emotion for a tweet, except when *no code* or *not relevant* were selected, in which case no additional options could be picked. The annotation of all the tweets was performed independently by three sociology PhD students. Out of the 3,759 tweets that were released for annotation, at least 2 of the annotators agreed in 3,085 cases (82.1%). We use the collection resulting from these 3,085 tweets as our target-domain dataset for classifier adaptation and evaluation. Note that tweets labelled as *no code* or *not relevant* are included in our dataset to reflect a more realistic data distribution on Twitter, while our source-domain data doesn't have any *no code* or *not relevant* tweets.

The distribution of emotion annotations in Table 2 shows a remarkable class imbalance, where *happy* accounts for 30.2% of the tweets, while the other emotions are seldom observed in the museum dataset. There is also a large number of tweets with no emotion associated (41.8%). One intuitive explanation is that Twitter users tend to express positive and appreciative emotions regarding their museum experiences and shy away from making negative comments. This can also be demonstrated by comparing the museum data emotion distribution to our general-domain source data as seen in Figure 1, where the sample ratio of positive instances is shown for each emotion category.

To quantify the difference between two text datasets, Kullback-Leibler (KL) divergence has been commonly used before [Dai *et al.*, 2007]. Here we use the KL-divergence method proposed by [Bigi, 2003], as it suggests a back-off smoothing method that deals with the data sparseness problem. Such back-off method keeps the probability distributions summing to 1 and allows operating on the entire vocabulary, by introducing a normalisation coefficient and a very small threshold probability for all the terms that are not in the given vocabulary. Since our source-domain data contains many more tweets than the target-domain data, we have randomly sub-sampled the former and made sure the two data sets have similar vocabulary size in order to avoid biases. We removed stop words, user mentions, URL links and re-tweet symbols prior to computing the KL-divergence. Finally we randomly split each data set into 10 folds and compute the in-domain and cross-domain symmetric KL-divergence (KLD) value between every pair of folds. Table 1 shows the computed KL-divergence averages. It can be seen that KL-divergence between the two data sets (i.e. $\text{KLD}(D_{src} \,||\, D_{tar})$) is twice as large as the in-domain KL-divergence values. This suggests a significant difference between data distributions in the two domain and thus justifies our need for domain adaptation.

Data domain	Averaged KLD value
$\text{KLD}(D_{src} \,||\, D_{src})$	2.391
$\text{KLD}(D_{tar} \,||\, D_{tar})$	2.165
$\text{KLD}(D_{src} \,||\, D_{tar})$	4.818

Table 1: In-domain and cross-domain KL-divergence values

Emotion	No. of tweets	% of tweets
no code	1572	41.8%
happy	1137	30.2%
not relevant	214	5.7%
anger	57	1.5%
surprise	35	0.9%
sad	32	0.9%
happy & surprise	11	0.3%
happy & sad	9	0.2%
disgust & anger	7	0.2%
disgust	6	0.2%
sad & anger	2	0.1%
sad & disgust	2	0.1%
sad & disgust & anger	1	<0.1%

Table 2: Target data emotion distribution

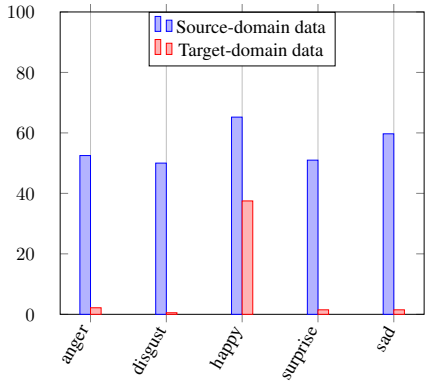


Figure 1: Source and target data distribution comparison

4 Methodology

Given the source-domain D_{src} and target-domain D_{tar}, we have one or k sets of labelled source-domain data denoted as $\left\{(x_i^k, y_i^k)\right\}_{i=1}^{N_{src}^k}$ in D_{src}, where x_i^k is the i_{th} feature vector with each element as the value of the corresponding feature and y_i^k are the emotion categories that the i_{th} instance belongs to. Suppose we have some classifiers $f_{src}^k(x)$ that have been trained on the source-domain data (named as the *auxiliary classifiers* in [Yang *et al.*, 2007]) and a small set of labelled target-domain data as D_{tar}^l where $D_{tar} \quad D_{tar}^l \cup D_{tar}^u$, our goal is to adapt $f_{src}^k(x)$ to a new classifier $f_{tar}(x)$ based on the small set of labelled examples in D_{tar}^l, so it can be used to accurately predict the emotion class of unseen data from D_{tar}^u.

4.1 Base Classifiers

Our base classifiers are the classifiers that have been trained on the source-domain data $\left\{(x_i, y_i)\right\}_{i=1}^{N_{src}}$, where $y_i \in \{1, ..., K\}$ with K referring to the number of emotion categories. In our work, we use Support Vector Machines (SVMs) in a "one-versus-all" setting, which trains K binary classifiers, each separating one class from the rest. We chose this as a better way of dealing with class imbalance in a multi-class scenario.

Features
The base classifiers are trained on 3 sets of features generated from the source-domain data: (i) n-grams, (ii) lexicon features, (iii) word embedding features.

N-gram models have long been used in NLP for various tasks. We used 1-2-3 grams after filtering out all the stop words, as our n-gram features. We construct 32 **Lexicon features** from 9 Twitter specific and general-purpose lexica. Each lexicon provides either a numeric sentiment score, or categories where a category could correspond to a particular emotion or a strong/weak positive/negative sentiment.

The use of **Word embedding features** to represent the context of words and concepts, has been shown to be very effective in boosting the performance of sentiment classification. In this work we use a set of word embeddings learnt using a sentiment-specific method in [Tang *et al.*, 2014] and another set of general word embeddings trained with 5 million tweets by [Vo and Zhang, 2015]. Training on an additional set of 3 million tweets we trained ourselves did not increase performance. Pooling functions are essential and particularly effective for feature selection from dense embedding feature vectors. [Tang *et al.*, 2014] applied the *max*, *min* and *mean* pooling functions and found them to be highly useful. We tested and evaluated six pooling functions, namely *sum*, *max*, *min*, *mean*, *std* (i.e. standard deviation) and *product*, and selected *sum*, *max* and *mean* as they led to the best performance.

4.2 Classifier Adaptation

[Yang *et al.*, 2007] proposes a many-to-one SVM adaptation model, which directly modifies the decision function of an ensemble of existing classifiers $f^k_{src}(x)$, trained with one or k sets of labelled source-domain data in D_{src}, and thus creates a new adapted classifier $f_{tar}(x)$ for the target-domain D_{tar}. The adapted classifier has the following form:

$$f_{tar}(x) \quad \sum_{k=1}^{M} \tau^k f^k_{src}(x) \qquad f(x) \tag{1}$$

where $\tau^k \in (0,1)$ is the weight of each base classifier $f^k_{src}(x)$. $f(x)$ is the perturbation function that is learnt from a small set of labelled target-domain data in D^l_{tar}. As shown in [Yang *et al.*, 2007] it has the form:

$$f(x) \quad w^T\phi(x) \quad \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{K}(x_i, x) \tag{2}$$

where $w \quad \sum_{i=1}^{N} \alpha_i y_i \phi(x_i)$ are the model parameters to be estimated from the labelled examples in D^l_{tar} and α_i is the feature coefficient of the i_{th} labelled target-domain instance. Furthermore $\boldsymbol{K}(\cdot,\cdot) \equiv \phi(\cdot)^T\phi(\cdot)$ is the kernel function induced from the nonlinear feature mapping. $f(x)$ is learnt in a framework that aims to minimise the regularised empirical risk [Yang, 2009]. The adapted classifier $f_{tar}(x)$ learnt under this framework tries to minimise the classification error on the labelled target-domain examples and the distance from the base classifiers $f^k_{src}(x)$, to achieve a better bias-variance trade-off.

In this work we use the extended multi-classifier adaptation framework proposed by [Yang and Hauptmann, 2008], which allows the weight controls $\{\tau^k\}_{k=1}^M$ of the base classifiers $f^k_{src}(x)$ to be learnt automatically based on their classification performance of the small set of labelled target-domain examples. To achieve this, [Yang and Hauptmann, 2008] adds another regulariser to the regularised loss minimisation framework, with the objective function of training the adaptive classifier now written as:

$$\begin{aligned} \min_{w,\tau,\xi} \quad & \frac{1}{2}w^T w \quad \frac{1}{2}B(\tau)^T\tau \quad C\sum_{i=1}^{N}\xi_i \\ \text{s.t.} \quad & y_i \sum_{k=1}^{M} \tau^k f^k_{src}(x) \quad y_i w^T\phi(x_i) \geq 1-\xi_i, \\ & \xi^m_i \geq 0, \forall (x_i, y_i) \in D_{src} \end{aligned} \tag{3}$$

where $\frac{1}{2}(\tau)^T\tau$ measures the overall contribution of base classifiers. Thus this objective function seeks to avoid over reliance on the base classifiers and also over-complex $f(\cdot)$. The two goals are balanced by the parameter B. By rewriting this objective function as a minimisation problem of a Lagrange (primal) function and set its derivative against w, τ, and ξ to zero, we have:

$$w \quad \sum_{i=1}^{N} \alpha_i y_i \phi(x_i), \quad \tau^k \quad \frac{1}{B}\sum_{i=1}^{N} \alpha_i y_i f^k_{src}(x_i) \tag{4}$$

where τ^k is a weighted sum of $y_i f^k_{src}(x_i)$ and it indicates the classification performance of f^k_{src} on the target-domain. Therefore we have base classifiers assigned with larger weight if they classify the labelled target-domain data well. Now given (1), (2) and (4), the new decision function can be formulated as:

$$\begin{aligned} f_{tar}(x) \quad & \frac{1}{B}\sum_{k=1}^{M}\sum_{i=1}^{N} \alpha_i y_i f^k_{src}(x_i) f^k_{src}(x) \qquad f(x) \\ & \sum_{i=1}^{N} \alpha_i y_i \Big(\boldsymbol{K}(x_i, x) \quad \frac{1}{B}\sum_{k=1}^{M} f^k_{src}(x_i) f^k_{src}(x) \Big) \end{aligned} \tag{5}$$

Comparing (5) with a standard SVM model $f(x)$ $\sum_{i=1} \alpha_i y_i \boldsymbol{K}(x_i, x)$, this multi-classifier adaptation model can be interpreted as a way of adding the predicted labels of base classifiers on the target-domain as additional features. Under this interpretation the scalar B balances the contribution of the original features and additional features.

4.3 Data Preprocessing

A set of preprocessing techniques applied include substituting URL links with strings "URL", user mentions with "@USERID", removing the hashtag symbol "#", normalising emoticons and abbreviations[3].

5 Results and Evaluation

In this section we present the experimental results and compare our proposed adaptation system with a set of domain-dependent and domain-independent strategies. We also investigate the effect of different sizes of the labelled target-domain data in the classification performance.

[3] `http://bit.ly/1U7fiQR`

5.1 Adaptation Baselines

The baseline methods and our proposed system are the following:

- **BASE**: the base classifiers use either one set of features or all three feature sets (i.e. BASE-all). As an example, the BASE-embedding classifier is trained and tuned with all source-domain data using only word-embedding features, then tested on 30% of our target-domain data. We use the LIBSVM implementation [Chang and Lin, 2011] of SVM for building the base classifiers.
- **TARG**: trained and tuned with 70% labelled target-domain data. Since this model is entirely trained from the target domain, it can be considered as the *performance upper-bound* that is very hard to beat.
- **AGGR**: an aggregate model trained from all source-domain data and 70% labelled target-domain data.
- **ENSEMBLE**: combines the base classifiers in an ensemble model. Then perform classification on 30% of the target-domain data to generate new training data, as described in Section 2.
- **ADAPT**: our domain adapted models using either one base classifier trained with all feature sets (i.e. ADAPT-1-model) or an ensemble of three standalone base classifiers with each trained with one set of features (i.e. ADAPT-3-model). We use 30% of the labelled target-domain data for classifier adaptation and parameter tuning described in Section 4.2.

The above methods are all tested on the same 30% labelled target-domain data in order to make their results comparable. In addition we perform in-domain cross-validation and evaluation only on our source-domain data using all feature sets; this model is named as SRC-all. We use an RBF kernel function (as it outperforms linear kernel. Polynomial kernel gives similar performance but requires more parameter tuning) with default setting of the gamma parameter γ in all the methods. For the cost factor C and class weight parameter (except the SRC-all model) we conduct cross-validated grid-search over the same set of parameter values for all the methods, for parameter optimisation. This makes sure our ADAPT models are comparable with BASE, TARG, ENSEMBLE and AGGR. For ADAPT-3-model we also optimise the base classifier weight parameters, denoted as τ^k in Eq.(1), as described in Section 4.2.

5.2 Experimental Results

We report the experimental results in **Table 3**, with three categories of models: 1) in-domain no adaptation methods, i.e. BASE and TARG models, TARG being the *upper-bound* for performance evaluation; 2) the domain adaptation baselines, i.e. AGGR and ENSEMBLE and 3) our adaptation systems (ADAPT models). As can be seen the classification performances reported for emotions other than "happy" are below 50 in terms of $\mathbf{F_1}$ score with some results being as low as 0.00. This is caused by the class imbalance issue within these emotions as shown in Table 2 and Figure 1, especially for the emotion "disgust" which has only 16 tweets. We tried to balance this issue using a class weight parameter, but it still is very challenging to overcome without acquiring more labelled data than we currently have. It especially effects our domain adaptation as all the parameters in Eq.(3) cannot be properly optimised.

Since there are very few tweets annotated as "disgust", we decide not to consider the "disgust" emotion as part of our experiment evaluation here. As seen in Table 3, BASE models are outperformed significantly by all other methods (except ENSEMBLE, which performs only slightly better than the BASE models) positing the importance of domain adaptation. With the exception of the ADAPT-3-model for "Anger", our ADAPT models consistently outperform AGGR-all and ENSEMBLE while showing competitive performance compared to the *upper-bound* baseline, TARG-all. We also observe that the aggregation model AGGR-all is outperformed by TARG-all, indicating such domain knowledge cannot be transferred effectively to a different domain by simply modelling from aggregated data from both domains. In comparison, our ADAPT models are able to leverage the large and balanced source-domain data (as base classifiers) unlike TARG, while adjusting the contribution of each base classifier unlike AGGR.

When comparing our ADAPT models, we find that in most cases models adapted from multiple base classifiers beat the ones adapted from one single base classifier, even though the same features are used in both scenarios. This shows the benefit of the multi-classifier adaptation approach, which aims to maximise the utility of each base classifier. Two additional models, namely ADAPT-1-modelx and ADAPT-3-modelx, are the replicates of ADAPT-1/3-model except they also use 40% target-domain data for tuning the model parameters. On average their results are only slightly better than ADAPT-1/3-model that use 30% of the target-domain data for both training and parameter optimisation. This is especially prominent with "happiness" where we have sufficient target-domain instances and less of a class imbalance issue. This shows our ADAPT models are able to yield knowledge transfer effectively across different domains with a small amount of labelled target-domain data. More analysis on the impact of adaptation sample ratios is given in Section 5.3.

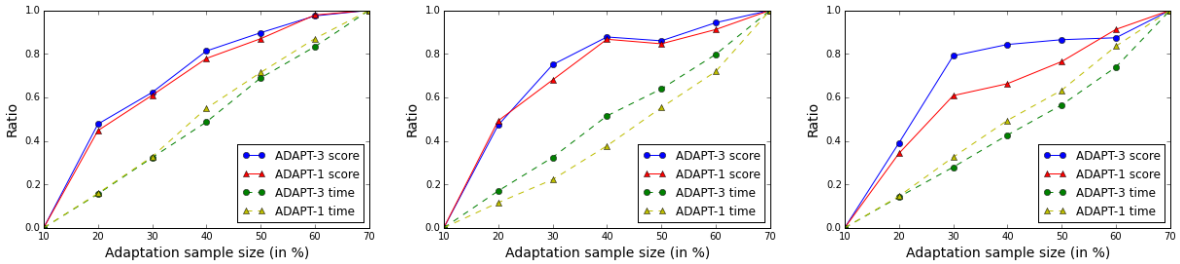
We can also evaluate the performance of each model by comparing its efficiency in terms of computation time. Here we report the total computation time taken for all the above methods except BASE, for the emotion "happiness". Such computation process consists of adaptation training, grid-search over the same set of parameter values and final testing. As seen in Table 4, compared to other out-of-domain strategies the proposed ADAPT models are more efficient to train especially in comparison with AGGR, which is an order of magnitude more costly due to the inclusion of source-domain data. Within the ADAPT models, ADAPT-1-model requires less time to train since it only has one base classifier for adaptation.

5.3 Effect of Adaptation Training Sample ratios

Here we evaluate the effect of different ratios of the labelled target-domain data on the overall classification performance for the emotion "happiness". Figure 2 shows the normalised $\mathbf{F_1}$ scores and computation time of each ADAPT

Model	Anger			Disgust			Happy			Surprise			Sad		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
BASE-ngrams	5.77	40.91	10.11	0.49	100.0	0.97	37.62	100.0	54.67	1.46	100.0	2.87	1.50	100.0	2.96
BASE-lexicon	2.59	90.91	5.03	0.55	100.0	1.10	38.43	98.96	55.36	0.00	0.00	0.00	2.54	93.33	4.94
BASE-embedding	2.06	72.73	4.02	0.00	0.00	0.00	39.18	96.11	55.66	2.00	60.00	3.88	1.49	80.00	2.92
BASE-all	2.01	59.09	3.88	5.00	20.00	8.00	38.75	98.19	55.57	1.69	66.67	3.29	1.58	86.67	3.11
TARG-all	36.00	40.91	**38.30**	0.00	0.00	0.00	78.04	84.72	81.24	20.83	33.33	**25.64**	18.75	20.00	**19.35**
AGGR-all	10.71	27.27	15.38	33.33	20.00	25.00	64.79	86.27	74.00	5.88	11.11	7.69	4.17	20.00	6.90
ENSEMBLE	2.11	100.0	4.13	0.49	100.0	0.97	45.20	83.55	58.66	2.70	11.11	4.35	1.46	100.0	2.88
ADAPT-1-model	16.28	31.82	21.54	0.59	80.00	1.18	79.34	80.57	79.95	11.11	13.33	12.12	100.0	6.67	12.50
ADAPT-3-model	20.00	9.09	12.50	0.00	0.00	0.00	82.11	80.83	81.46	8.14	46.67	13.86	8.77	33.33	13.89
ADAPT-1-modelx	21.43	13.64	16.67	100.0	20.00	**33.33**	80.53	79.27	79.90	12.50	26.67	17.02	16.67	13.33	14.81
ADAPT-3-modelx	20.00	22.73	21.28	1.82	20.00	3.33	80.30	83.42	**81.83**	12.50	26.67	17.02	10.20	33.33	15.63
SRC-all	93.57	93.37	93.46	99.05	98.73	98.89	81.87	85.91	83.83	96.25	98.03	97.13	91.04	92.51	91.76

Table 3: Model performance comparison



(a) $C = 1$ (b) $C = 3$ (c) $C = 10$

Figure 2: Performance of each ADAPT model with C = 1,3,10 vs. its computation time

Model	Total computation time in minutes
TARG-all	7.72
ENSEMBLE	209.72
AGGR-all	1238.24
ADAPT-1-model	26.30
ADAPT-3-model	118.41

Table 4: Total computation time for each method

model across different adaptation training sample sizes ranging from 10% to 70% of the total target-domain data (with the same 30% held out as test data) and with the cost factor C = 1, 3 and 10 (as the same choices of C are used in [Yang *et al.*, 2007] for conducting their experiment). We observe a logarithmic growth for the $\mathbf{F_1}$ scores obtained from every model, against a linear growth of computation time cost. Thus even though there is a reasonable increase in classification performance when increasing the adaptation sample size from 50% to 70%, it becomes much less efficient to train such models and we require more data, which may not be available. Since we have a trade-off between model effectiveness and efficiency here, it is appropriate to use 30% of our labelled target-domain data for classifier adaptation as we have done so in ADAPT-1-model and ADAPT-3-model. One should select the adaptation training sample size accordingly based on the test data at hand, but empirically we think 1,000 labelled target-domain tweets would be enough for an effective adaptation to classify 3,000-4,000 test tweets.

6 Conclusion

In this work we study a model-based multi-class adaptive-SVM approach to cross-domain emotion recognition and compare against a set of domain-dependent and domain-independent strategies. We conduct a series of experiments and evaluate our proposed system on a set of newly annotated Twitter data about museums. We find that our adapted SVM model outperforms the out-of-domain base models and domain adaptation baselines while also showing competitive performance against the in-domain *upper-bound* model. Moreover, in comparison to other adaptation strategies our approach is computationally more efficient especially compared to the classifier trained on aggregated source and target data. Finally, we shed light on how different ratios of labelled target-domain data used for adaptation can effect classification performance. We show there is a trade-off between model effectiveness and efficiency when selecting adaptation sample size. Our code and data[4] are publicly available, enabling further research and comparison with our approach.

In the future we would like to investigate a feature-based deep learning approach for cross-topic emotion classification on Twitter while examining the possibility of making it as efficient and flexible as the model adaptation based approaches. Another future direction is to study how to best resolve the remarkable class imbalance issue in social media emotion analysis when some emotions are rarely expressed.

[4] `http://bit.ly/1SddvIw`

Acknowledgments

This work has been funded by the AHRC SMILES project. We would like to thank Liz Walker, Matt Jeffryes and Michael Clapham for their contribution to earlier versions of the emotion classifiers.

References

[Bigi, 2003] Brigitte Bigi. *Using Kullback-Leibler distance for text categorization*. Springer, 2003.

[Blitzer *et al.*, 2007] John Blitzer, Mark Dredze, Fernando Pereira, et al. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, volume 7, pages 440–447, 2007.

[Bollegala *et al.*, 2011] Danushka Bollegala, David Weir, and John Carroll. Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In *NAACL HLT*, pages 132–141. Association for Computational Linguistics, 2011.

[Bollen *et al.*, 2011] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[Dai *et al.*, 2007] Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. Co-clustering based classification for out-of-domain documents. In *SIGKDD*, pages 210–219. ACM, 2007.

[Drotner and Schrøder, 2014] Kirsten Drotner and Kim Christian Schrøder. *Museum communication and social media: The connected museum*. Routledge, 2014.

[Feldman, 2013] Ronen Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[Fletcher and Lee, 2012] Adrienne Fletcher and Moon J Lee. Current social media uses and evaluations in american museums. *Museum Management and Curatorship*, 27(5):505–521, 2012.

[Glorot *et al.*, 2011] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, pages 513–520, 2011.

[Jiang and Zhai, 2007] Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *ACL*, pages 264–271. Association for Computational Linguistics, June 2007.

[Liu *et al.*, 2013] Shenghua Liu, Fuxin Li, Fangtao Li, Xueqi Cheng, and Huawei Shen. Adaptive co-training svm for sentiment classification on tweets. In *CIKM*, pages 2079–2088. ACM, 2013.

[Pan *et al.*, 2010] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In *WWW*, pages 751–760. ACM, 2010.

[Pang and Lee, 2008] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[Purver and Battersby, 2012] Matthew Purver and Stuart Battersby. Experimenting with distant supervision for emotion classification. In *EACL*, pages 482–491. Association for Computational Linguistics, 2012.

[Tang *et al.*, 2014] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for twitter sentiment classification. In *ACL*, volume 1, pages 1555–1565, 2014.

[Townsend *et al.*, 2014] Richard Townsend, Aaron Kalair, Ojas Kulkarni, Rob Procter, and Maria Liakata. University of warwick: Sentiadaptron-a domain adaptable sentiment analyser for tweets-meets semeval. *SemEval 2014*, page 768, 2014.

[Tsakalidis *et al.*, 2014] Adam Tsakalidis, Symeon Papadopoulos, and Ioannis Kompatsiaris. An ensemble model for cross-domain polarity classification on twitter. In *WISE*, pages 168–177. Springer, 2014.

[Tumasjan *et al.*, 2010] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

[Villaespesa, 2013] Elena Villaespesa. Diving into the museums social media stream: Analysis of the visitor experience in 140 characters. In *Museums and the Web*, 2013.

[Vo and Zhang, 2015] Duy-Tin Vo and Yue Zhang. Target-dependent twitter sentiment classification with rich automatic features. In *IJCAI*, pages 1347–1353, 2015.

[Xia *et al.*, 2013] Rui Xia, Chengqing Zong, Xuelei Hu, and Erik Cambria. Feature ensemble plus sample selection: domain adaptation for sentiment classification. *Intelligent Systems, IEEE*, 28(3):10–18, 2013.

[Xia *et al.*, 2014] Rui Xia, Jianfei Yu, Feng Xu, and Shumei Wang. Instance-based domain adaptation in nlp via in-target-domain logistic approximation. In *AAAI*, 2014.

[Yang and Hauptmann, 2008] Jun Yang and Alexander G Hauptmann. A framework for classifier adaptation and its applications in concept detection. In *MIR*, pages 467–474. ACM, 2008.

[Yang *et al.*, 2007] Jun Yang, Rong Yan, and Alexander G Hauptmann. Cross-domain video concept detection using adaptive svms. In *Proceedings of the 15th international conference on Multimedia*, pages 188–197. ACM, 2007.

[Yang, 2009] Jun Yang. *A general framework for classifier adaptation and its applications in multimedia*. PhD thesis, Columbia University, 2009.

Personality Trait Classification of Essays with the Application of Feature Reduction

Edward P. Tighe, Jennifer C. Ureta, Bernard Andrei L. Pollo,
Charibeth K. Cheng, and **Remedios de Dios Bulos**
De La Salle Univeristy, Manila, Philippines
{edward_p_tighe, jennifer_ureta, bernard_pollo}@dlsu.edu.ph,
chari.cheng@delasalle.ph, remedios.bulos@dlsu.edu.ph

Abstract

Determining an individual's personality traits is an important concept in Psychology. Although traits are normally assessed through self-report tests, an alternative method would be to computationally analyze an individual's linguistic markers. Studies in personality trait classification show promising results and look to continuously improve the field by either using new features or by collecting new data from social media; however, a key concept that is not always considered is the use of feature reduction techniques. This research aims to perform feature reduction techniques on linguistic features from essays and classify the author's personality traits based on the reduced feature set. The classifiers are evaluated by comparing against a baseline classifier trained with all extracted features. The feature reduction techniques used are Information Gain and Principal Component Analysis. The results show that feature reduction techniques are able to increase classification measures, but not by significant values. Reduced datasets are exceptionally beneficial in reducing the amount of data needed allowing classifiers to perform faster while still maintaining classification measures.

1 Introduction

Personality Psychology, or simply Personality, is "the scientific study of psychological forces that make people uniquely themselves" [Friedman and Schustack, 2014, p.1]. These forces consist of organized and relatively enduring traits and mechanisms that influence one's interactions with the intrapsychic, physical, and social environments [Larsen and Buss, 2008, p.4].

One of the most well-researched theories describing personality trait variation would be the Five Factor model, also known as the Big Five [Norman, 1963; Goldberg, 1981]. The Big Five is an organization of personality facets that are subsets of five broad traits: *Extraversion*, *Agreeableness*, *Conscientiousness*, *Neuroticism*, and *Openness to Experience*. John et al. [2008, p.116] mentions that these five traits "were derived from analyzing terms people use to describe themselves and others." The Big Five have been used in studies observing individuals in environments such as at work [Richardson *et al.*, 2009] and in academics [Komarraju *et al.*, 2009]. Research using both natural language adjectives and theoretically based personality instruments supports the comprehensiveness of the model and its applicability across observers and cultures [McCrae and John, 1992].

Personality traits are traditionally measured through the use of questionnaires such as the Big Five Inventory (BFI) [John *et al.*, 1991]; however, an alternative approach would be to analyze an individual's linguistic markers. An individual's choice of words eventually becomes consistent over time and context and can be used as an individual difference measure [Pennebaker *et al.*, 2003]. A study by Pennebaker and King [1999] showed multiple correlations between linguistic markers and the Big Five such as how Neuroticism is positively correlated with the use of negative emotion words and negatively with positive emotion words. Goldberg [1981, p.142] mentions that "the more important an individual difference [is] in human transactions; the more likely languages will have a term for it."

Correlations between linguistic markers and personality traits have paved the way for research in the area of automatic personality classification. One of the earliest studies [Mairesse *et al.*, 2007] focused on classifying personality traits based on text. They extracted linguistic features from essays using a text analysis tool and a psycholinguistic database. Their findings were modest, but still showed that computationally modeling the Big Five was possible. Subsequent studies were able to present promising methods in improving upon the findings of Mairesse et al. [2007] by introducing new linguistic features [Mohammad and Kiritchenko, 2013; Poria *et al.*, 2013a]. Other studies [Golbeck *et al.*, 2011; Schwartz *et al.*, 2013; Park *et al.*, 2014; Peng *et al.*, 2015] focus on different data sources by taking advantage of social media, and collecting or using data from users of these growing platforms.

Although there have been advancements in the field of personality trait classification, there are still gaps in determining which linguistic features are most significant for the classification process. This research investigates the use of two feature reduction techniques in order to improve the computation involved. The data source for this research is the Pennebaker and King [1999] dataset of essays. Features are extracted us-

Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016, pages 22-28, New York City, USA, July 10, 2016.

ing LIWC and are analyzed to see if they can be reduced to a smaller set while still being able to aid in classifying the Big Five. The classifiers are then built using the reduced set of features and are compared against classifiers using the complete set of features. This research aims to show that the use of feature reduction techniques are beneficial to future work in the field.

The remainder of this paper is organized as follows: Section 2 reviews studies on personality trait classification that work with the Pennebaker and King [1999] essay dataset. Section 3 discusses the characteristics of the data source used in this research. Section 4 explains how features are extracted from the data source. Section 5 explains the background of the feature reduction techniques used. Section 6 presents how each classifier was built. Section 7 discusses the overall performance of the classifiers and the effects of feature reduction. Finally, Section 8 concludes the paper and explains recommendations for future work.

2 Related Works

Personality Trait Classification based on linguistic markers is a growing field. Although other studies [Golbeck *et al.*, 2011; Schwartz *et al.*, 2013; Park *et al.*, 2014; Peng *et al.*, 2015] use big collections of data from social media, this paper limits its review to studies that use the Pennebaker and King [1999] essay dataset for the sake of having a common ground.

One of the earliest studies regarding automatic personality classification is that of Mairesse et al. [2007]. Their methods of extracting features relied on LIWC[1] [Pennebaker *et al.*, 2001] and the MRC Psycholinguistic Database [Coltheart, 1981]. LIWC produced a total of 88 features and is further discussed in Section 4. They also used 14 psycholinguistic features[2] from the MRC Psycholinguistic Database, a machine usable dictionary. They then trained classifiers based on different combinations of the set of features and had promising results. Openness to Experience was the easiest to identify among the Big Five having an accuracy of 62.5% using only LIWC features. They also showed how features from LIWC out performed those from MRC; however, both showed promising correlations to the Big Five. Their results were modest, but were significant enough to show that computationally modeling personality traits was possible.

One recent study [Mohammad and Kiritchenko, 2013] made use of fine affect or emotion category features as alternatives for personality trait classification. They were able to extract an extensive amount of emotion features with the use of the NCR Hashtag Emotion Lexicion [Mohammad and Turney, 2010]. This lexicon is able to produce either 8 basic emotions or 585 fine emotion features. They also made use of the Specificity Lexicon and Osgood Dimensions Lexicon [Turney and Littman, 2003]. The first lexicon calculated the average information content of an essay while the later was able to extract the average evaluativeness, activity, and potency scores of words. Finally, they made use of the LIWC features and frequencies of unigrams of the essays. They experimented by combining different features sets and ran them through Support Vector Machine classifiers. The classifier that performed best was built using the LIWC and the 585 fine emotion features. Their results showed minimal improvement over the results of Mairesse et al. [2007], but revealed that emotion category features contain information regarding an individual‘s personality and can be considered useful for future studies.

Another study [Poria *et al.*, 2013a] introduced a novelty approach of using of common sense knowledge. They utilize ConceptNet [Havasi *et al.*, 2007] and EmoSenticNet [Poria *et al.*, 2013b] to extract sentiment polarity scores and affective labels from the essays. They also extract linguistic features from LIWC and MRC. They train Support Vector Machine classifiers and compare against Mairesse et al [Mairesse *et al.*, 2007] and Mohammad and Kiritchenko [Mohammad and Kiritchenko, 2013]. Their results show significant improvements demonstrating that the sentiment polarity and affective labels contained relevant information in classifying personality traits.

With the discovery of more and more features with information pertaining to an individual‘s personality, the issue of irrelevant features and overfitting arises. Each of the previously reviewed studies presents an opportunity to investigate the use of feature reduction due to the high volume of features presented. This research explores the application of feature reduction on LIWC features and aims to showcase the benefits of using these techniques.

3 Data Source

The data used in this research was gathered and used in a study by Pennebaker and King [1999]. The actual file was retrieved from myPersonality[3]. It consists of a total of 2,468 essays or daily writing submissions from 34 psychology students. There are a total of 29 women and 5 men whose ages ranged from 18 to 67 with a mean of 26.4 and a standard deviation of 11.1.

The writing submissions were in the form of a course requirement or assignment but were not graded. For each assignment, students were expected to write a minimum of 20 minutes per day about a specific topic. The data was collected during a 2-week summer course between 1993 to 1996. Each student completed their daily writing for 10 consecutive days.

Students' personality scores were assessed by answering the Big Five Inventory (BFI) [John *et al.*, 1991]. The BFI is a 44-item self-report questionnaire that provides a score for each of the five personality traits. Each item consists of short phrases and is rated using a 5-point scale that ranges from 1 (disagree strongly) to 5 (agree strongly).

An instance in the data source consists of a filename or ID, the actual essay, and five classification labels of the Big Five personality traits. Labels were originally in the form of either yes (‘y’) or no (‘n’) to indicate scoring high or low for a given trait; however, this research changed the labels to ‘y’ to 1 and ‘n’ to 0 according to the preference of the researchers.

[1] It was assumed by the researchers that Mairesse et al. [2007] used the 2001 version of LIWC as to how it was cited in their paper

[2] However, the total number of features is listed as 26 [Wilson, 1988]

[3] www.mypersonality.org

4 Feature Extraction

In order to extract information from raw text, LIWC[4] was utilized. LIWC stands for Linguistic Inquiry and Word Count and was developed by Pennebaker et al. [2007]. It is a text analysis tool that provides an efficient and effective method for studying the various emotional, cognitive, and structural components present in individuals' written samples.

The tool analyses text files sequentially, one target word at a time by searching through it dictionary file. If the target word matches the dictionary word, the appropriate word category scale is incremented. Pennebaker et al. [2007] explains that there are a total of 80 output features consisting of 4 general descriptor categories (e.g., total word count, words per sentence), 22 standard linguistic dimensions (e.g., frequency of pronouns, articles), 32 word categories tapping psychological constructs (e.g., affect, cognition), 7 personal concern categories (e.g., work, home), 3 paralinguistic dimensions (assents, fillers, nonfluencies), and 12 punctuation categories (e.g., periods, commas). Values of all features except word count and words per sentence reflect percentage of total words.

Based on the methods of Mairesse et al. [2007], the essays from the data source were fed through LIWC. The output features were used to create a dataset for each of the Big Five. These datasets contain all 80 LIWC features and one of the five personality trait as the classification label. It is important to note that replication of the methods used by Mairesse et al. [2007] was chosen by the researchers as the better alternative to direct comparison of results. This was due to the difference in the number of output features from the version of LIWC reported in their paper. Replication of methods would allow for a better baseline when comparing against a reduced set of features.

5 Feature Reduction

LIWC provides a vast amount of information that it would be important to analyze whether or not classification of the Big Five can be improved by reducing the set of features. The presence of non-relevant features can influence a classifier to produce smaller error by fitting the model according to the training data. Removing such features can increase the predictive power of a classifier by focusing only on certain features. A defined model may be able to classify unseen data better and is desirable for real world scenarios. Feature reduction becomes an important concept to consider when trying to improve classification. The techniques that are performed are Information Gain and Principal Component Analysis.

The following subsections discuss how these techniques work and present their respective output. Both techniques were applied on the datasets using Waikato Environment for Knowledge Analysis or Weka, a tool of machine learning algorithms and data preprocessing [Hall *et al.*, 2009].

5.1 Information Gain

Information Gain is a measure of how effective a given feature is in classifying data [Mitchell, 1997]. It becomes essential to this research to evaluate each of the 80 LIWC features and determine which provide significance in classifying the Big Five.

One important concept in computing for the Information Gain is Entropy or being able to characterizes the impurity of an arbitrary collection of examples. Entropy is defined as

$$E(F) \quad -\sum_{v=1}^{n} p_v \log_2 p_v$$

where F is a feature or classification containing a number n of different discrete values and where p_v is the proportion of F belonging to value v. When values are continuous in nature, the values are discretized by splitting at a point that provides the maximum information gain. Therefore, the information gain of a feature F relative to a classification C can be defined as

$$IG(C,F) \quad E(C) - \sum_{v \in Values(F)} \frac{|C_v|}{|C|} E(C_v)$$

where $Values(F)$ is the set of all possible values for feature A, and C_v is the subset of C for which feature F has values v. Basically, information gain is the entropy of class C reduced by the weighted average entropy of each subset S_v.

The Information Gain of all 80 features were computed for each of the Big Five datasets. For each of the datasets, only features with non-zero information gain were selected. Table 1 shows the remaining features per personality trait along with their respective information gain.

5.2 Principal Component Analysis

Principal Components Analysis (PCA) is used to identify patterns, and highlight the similarities and differences in data [Smith, 2002]. It is particularly useful when dealing with data with a high number of features as it is able to reduce the number of these features without losing much information. Concepts of covariance, matrix operations, eigenvalues, and eigenvectors are all used to compute for the principal components.

Smith [2002] explains that in order to perform PCA, the first step would be to calculate all the features' covariance matrix CM which can be defined for a set of data with m features as

$$CM^{m \times m} \quad \begin{pmatrix} cov(F_1,F_1) & \cdots & cov(F_1,F_m) \\ \vdots & \ddots & \vdots \\ cov(F_m,F_1) & \cdots & cov(F_m,F_m) \end{pmatrix}$$

where $CM^{m \times m}$ is a matrix with m rows and m columns composed the covariance between features Fx where x ranges from 1 to m. The second step is to calculate the CM's eigenvectors and their respective eigenvalue. Once found, the eigenvalues are ranked from highest to lowest and are removed along with their paired eigenvector according to a set threshold. The remaining eigenvectors are then inserted into a Feature Vector FV from highest to lowest eigenvalue. The final dataset values $FinalData$ is defined as

$$FinalData \quad FV^{\mathsf{T}} \times AdjustedValues^{\mathsf{T}}$$

[4] Developed in 2007 and is a different version than the previously mentioned LIWC in Section 2

Table 1: The remaining features for each of the Big Five after removing LIWC features with zero information gain

Extraversion		***Conscientiousness***			
Gain	*Feature*	*Gain*	*Feature*	*Gain*	*Feature*
0.00696	Articles	0.00996	Swear words	0.00688	Apostrophes
0.00643	Personal Pronouns	0.00975	Anger	0.00684	Function Words
0.00637	Sexuality	0.00825	Negative Emotion	0.00659	Prepositions
0.00589	Conjunctions	0.00731	Dictionary Words	0.00659	Exclamation Marks
Openness to Experience					
Gain	*Feature*	*Gain*	*Feature*	*Gain*	*Feature*
0.01583	Dictionary Words	0.00836	Friends	0.00728	Personal Pronouns
0.01496	Work	0.00808	All Punctuation	0.00712	Cognitive Processes
0.01300	First Person Singular Pronoun	0.00807	Motion	0.00661	Function Words
0.01216	Second Person Pronoun	0.00802	Religion	0.00655	Words $>$ 6 Letters
0.01046	Home	0.00763	Parentheses	0.00644	Question Marks
0.00978	Time	0.00760	Articles	0.00623	Sexuality
0.00937	Relativity	0.00736	Commas	0.00552	Quotation Marks
0.00844	Swear Words	0.00735	Regular Verbs	0.00534	Anger
Agreeableness		***Neuroticism***			
Gain	*Feature*	*Gain*	*Feature*	*Gain*	*Feature*
0.01103	Anger	0.01948	Negative Emotion	0.00647	Dictionary Words
0.00987	Swear Words	0.00911	Sadness	0.00635	Total Pronouns
0.00708	Negative Emotion	0.00895	First Person Singular Pronoun	0.00595	Negations
0.00680	Family	0.00750	Anxiety	0.00555	Leisure
0.00647	Dictionary Words	0.00744	Personal Pronoun		

where the transposed FV is multipled with a transposed matrix $AdjustedValues$ containing the original datasets's values adjusted by each feature's mean.

After performing Weka's implementation of PCA on all 80 LIWC features, a total of 56 eigenvectors were found and used to create a new dataset for each of the Big Five.

6 Classification

A 10-fold cross validation was performed on each of the 15 datasets (5 using all features, 5 using information gain, and 5 using PCA) in order to evaluate their overall effectiveness. This research recorded the accuracy, precision, and F-measure of all classifiers and the amount of reduction in terms of a dataset's feature size.

Each dataset was fed through three learning algorithms in Weka and compared against a baseline classifier (*ZeroR*) that returned the majority class. The algorithms used are two implementations of Support Vector Machine (*libSVM* and *SMO*), and Linear Logistic Regression (*SimpleLogistic*). Other algorithms such as k-Nearest Neighbour (*IBk*, where k equaled *1* and *5*), C4.5 Decision Tree (*J48*), Naive Bayes (*NaiveBayes*), and Random Forest (*RandomForest*) were also investigate; however, these classifiers were discarded due to poor performance. Default parameter settings were used for each of the learning algorithms.

7 Discussion

An overview of the results of classification, as seen in Table 2, shows that Openness to Experience is the easiest trait to identify regardless of feature reduction techniques. The remaining traits, ranked from easiest to hardest to identify, are Neuroticism, Agreeableness, Conscientiousness, and Extraversion. The ranking corresponds to the Information Gain of each LIWC feature per Big Five as seen in Table 1. Openness to Experience had the most amount of features remaining after removing those with zero information gain. On the other hand, Extraversion had the least remaining features. Each of the remaining features can also serve as a descriptor of how word choice of an individual is related to their personality traits. The use of *Negative Emotions* is relevant in determining one's Conscientiousness, Agreeableness, and Neuroticism. Similarly, *Swear Words* is relevant to Conscientiousness, Openness to Experience and Agreeableness. Although, it is important to note that that higher usage rate of a linguistic feature does not equates to a high personality trait score. Features with higher Information Gain simply indicate that a feature is more effective in separating and classifying data.

The classifiers using feature reduced datasets were generally able to increase classification measures, but not by significant values. This suggests that the LIWC features do not have any more information to provide in classifying the Big Five, at least when considering only a feature set of only LIWC features. A better comparison of the classifiers built using all 80 LIWC features and the feature reduced datasets is shown in Table 3. This indicates that classifiers using feature reduced datasets were able to slightly edge out classifiers using all features in four of the five personality traits. Agreeableness was the only trait where both best classifiers performed similarly.

This research also noted the minimal increase in classi-

Table 2: The performance of each classifier according to their respective dataset

Personality Trait	Features Used	Number of Feature	Size Reduction	Classifier	Accuracy	Precision	F-measure
Agreeableness	All Features	80	0.00%	ZeroR	53.08%	0.282	0.368
				LibSVM	51.78%	0.502*	0.475*
				SMO	56.20%	0.559*	0.549*
				SimpleLogistic	57.42%*	0.572*	0.566*
	Information Gain	5	93.75%	LibSVM	55.47%	0.551	0.536**
				SMO	55.88%	0.578	0.487
				SimpleLogistic	57.54%	0.575	0.557
	PCA	56	30.00%	LibSVM	57.05%**	0.568**	0.563**
				SMO	55.15%	0.547	0.535
				SimpleLogistic	55.55%	0.552	0.547
Conscientiousness	All Features	80	00.00%	ZeroR	50.81%	0.258	0.342
				LibSVM	51.99%	0.520*	0.520*
				SMO	54.82%	0.548*	0.545*
				SimpleLogistic	54.91%*	0.549*	0.548*
	Information Gain	8	90.00%	LibSVM	53.73%	0.537	0.537
				SMO	55.55%	0.560	0.541
				SimpleLogistic	55.27%	0.553	0.550
	PCA	56	30.00%	LibSVM	56.04%**	0.560**	0.560**
				SMO	55.80%	0.558	0.554
				SimpleLogistic	55.39%	0.554	0.552
Extraversion	All Features	80	00.00%	ZeroR	51.74%	0.268	0.353
				LibSVM	51.22%	0.507*	0.492*
				SMO	53.85%	0.537*	0.533*
				SimpleLogistic	53.49%	0.533*	0.530*
	Information Gain	4	95.00%	LibSVM	52.82%	0.527	0.518
				SMO	54.74%	0.547	0.532
				SimpleLogistic	54.70%	0.546	0.541
	PCA	56	30.00%	LibSVM	55.75%**	0.557**	0.556**
				SMO	53.65%	0.535	0.531
				SimpleLogistic	53.00%	0.528	0.527
Neuroticism	All Features	80	00.00%	ZeroR	50.04%	0.250	0.334
				LibSVM	51.09%	0.511*	0.509*
				SMO	57.05%	0.571*	0.570*
				SimpleLogistic	57.46%*	0.575*	0.575*
	Information Gain	9	88.75%	LibSVM	55.79%**	0.558**	0.558**
				SMO	57.13%	0.572	0.570
				SimpleLogistic	57.45%	0.575	0.574
	PCA	56	30.00%	LibSVM	58.31%**	0.583**	0.583**
				SMO	56.69%	0.567	0.567
				SimpleLogistic	57.17%	0.572	0.572
Openness to Experience	All Features	80	00.00%	ZeroR	51.54%	0.266	0.351
				LibSVM	54.05%	0.540*	0.524*
				SMO	61.26%*	0.613*	0.613*
				SimpleLogistic	59.52%*	0.595*	0.595*
	Information Gain	24	70.00%	LibSVM	56.93%	0.569	0.567
				SMO	61.83%	0.618	0.618
				SimpleLogistic	61.06%	0.610	0.610
	PCA	56	30.00%	LibSVM	59.92%	0.599	0.598
				SMO	61.95%	0.619	0.619
				SimpleLogistic	61.83%	0.618	0.618

Paired t-test was performed in Weka where significance was set at 0.05

*Significantly better than the baseline (*ZeroR*); **Significantly better than its respective classifier trained using all features

Table 3: Comparison of best performing classifiers using all features and using feature-reduced datasets

	Using All Features				Using Feature-Reduced Datasets			
Big Five	**Classifier**	**Accuracy**	**Precision**	**F-measure**	**Classifier**	**Accuracy**	**Precision**	**F-measure**
Agreeableness	SimpleLogistic	57.42%	0.572	0.566	SimpleLogistic[A]	57.54%	0.575	0.557
Conscientiousness	SimpleLogistic	54.91%	0.549	0.548	LibSVM[B]	56.04%	0.560	0.560
Extraversion	SMO	53.85%	0.537	0.533	LibSVM[B]	55.75%	0.557	0.556
Neuroticism	SimpleLogistic	57.46%	0.575	0.575	LibSVM[B]	58.31%	0.583	0.583
Openness to Experience	SMO	61.26%	0.613	0.613	SMO[B]	61.95%	0.619	0.619

Classifiers with [A] were trained using Information Gain reduced feature sets and [B] represents classifiers trained using PCA reduced feature sets

fication measures from using all features to using reduced features when dealing with Agreeableness. This can be attributed to how the classifier *SimpleLogistic* works. It is important to note that *SimpleLogistic* includes its own implementation of feature reduction [Landwehr *et al.*, 2005]. Interestingly, the only common attributes in comparison to those selected using Information Gain were *Anger* and *Family*. The remaining attributes selected by *SimpleLogistic* were *Words greater than 6 letters*, *Common Adverbs*, *Negations*, *Anxiety*, *Motion*, *Exclamation Marks*, and *Dashes*. The classifier also selected features for other personality traits that were not present in the set of remaining features after performing Information Gain.

Despite the results that feature reduction techniques were not able to significantly increase classification measures, it is important to note that the amount of reduction made in the size of the set of features is significant as seen in column *Size Reduction* of Table 2. Datasets using Information Gain had highly significant size reductions ranging from *70%* to *95%* while still able to perform up to par with those using all 80 LIWC feature. On the other hand, PCA was able to significantly reduce the set of features by *30%* while still covering *95%* of the feature set's variance. The PCA reduced dataset was also able to significantly improve the classifier *LibSVM* for all personality traits which is evident when looking at the measures. Agreeableness was able to improve from an F-measure of 0.475 to 0.563 resulting in a 0.088 increase. On the other hand, Conscientiousness had improved from an F-measure of 0.52 to 0.56 resulting in the the lowest increase of 0.04. As a whole, feature reduction techniques were able to improve the classification of personality traits by both slightly increasing classification measures and heavily reducing the size of the datasets.

8 Conclusion and Recommendations

This research was able to demonstrate that feature reduction techniques like Information Gain and Principal Component Analysis are beneficial in classifying an individual's personality traits based on text data. Applying these techniques reduced the size of the original data while slightly improving the classifiers' level of performance. A reduced-dataset leads to a more defined model which can better handle unseen data. This research was also able to highlight LIWC features that contain the most Information Gain about an individual's traits. This knowledge can be useful outside of computational classification by providing additional linguistic descriptors of individual's with certain personality traits.

This research also recommends two areas for improvements regarding future work in the field of text-based personality trait classification. The first would concern the data source containing binomially labeled traits and thereby categorizing an individual into one or the other. This representation does not capture the dimensional nature of a trait and would be better represented as either the raw output of a certain personality inventory or its normalized form. The second area for improvement would involve studying the use of non-western personality trait theories or indigenous measures. Such works are normally gauged towards understand a particular culture and would be a good area to apply linguistic analysis. Findings would be beneficial to both culture-specific and cross-cultural psychology.

References

[Coltheart, 1981] Max Coltheart. The MRC Psycholinguistic Database. *The Quarterly Journal of Experimental Psychology*, 33(4):497–505, 1981.

[Friedman and Schustack, 2014] Howard S. Friedman and Miriam W. Schustack. *Personality: Classic Theories and Modern Research: Pearson New International Edition*. Pearson Education Limited, 2014.

[Golbeck *et al.*, 2011] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. Predicting Personality from Twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 149–156. IEEE, 2011.

[Goldberg, 1981] Lewis R. Goldberg. Language and Individual Differences: The Search for Universals in Personality Lexicons. *Review of Personality and Social Psychology*, 2(1):141–165, 1981.

[Hall *et al.*, 2009] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[Havasi *et al.*, 2007] Catherine Havasi, Robert Speer, and Jason Alonso. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent advances in natural language processing*, pages 27–29. Citeseer, 2007.

[John *et al.*, 1991] Oliver P. John, Eileen M. Donahue, and Robert L. Kentle. The Big Five Inventory–Versions 4a and 54. Berkeley, CA: Univeristy of California, Berkeley, Institute of Personality and Social Research, 1991.

[John *et al.*, 2008] Oliver P. John, Laura P. Naumann, and Christopher J. Soto. Paradigm Shift to the Integrative Big Five Trait Taxonomy. *Handbook of Personality: Theory and Research*, 3:114–158, 2008.

[Komarraju *et al.*, 2009] Meera Komarraju, Steven J. Karau, and Ronald R Schmeck. Role of the Big Five Personality Traits in Predicting College Students' Academic Motivation and Achievement. *Learning and Individual Differences*, 19(1):47–52, 2009.

[Landwehr *et al.*, 2005] Niels Landwehr, Mark Hall, and Eibe Frank. Logistic model trees. *Machine Learning*, 59(1-2):161–205, 2005.

[Larsen and Buss, 2008] Randy J. Larsen and David M. Buss. *Personality Psychology: Domains of Knowledge About Human Nature*. McGraw Hill, 2008.

[Mairesse *et al.*, 2007] François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, pages 457–500, 2007.

[McCrae and John, 1992] Robert R. McCrae and Oliver P. John. An Introduction to the Five-Factor Model and its Applications. *Journal of Personality*, 60(2):175–215, 1992.

[Mitchell, 1997] Tom M. Mitchell. *Machine learning. WCB*. McGraw-Hill Boston, MA:, 1997.

[Mohammad and Kiritchenko, 2013] Saif M. Mohammad and Svetlana Kiritchenko. Using Nuances of Emotion to Identify Personality. *arXiv preprint arXiv:1309.6352*, 2013.

[Mohammad and Turney, 2010] Saif M Mohammad and Peter D Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics, 2010.

[Norman, 1963] Warren T. Norman. Toward an Adequate Taxonomy of Personality Attributes: Replicated Factor Structure in Peer Nomination Personality Ratings. *The Journal of Abnormal and Social Psychology*, 66(6):574, 1963.

[Park *et al.*, 2014] Gregory Park, H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Michal Kosinski, David J. Stillwell, Lyle H. Ungar, and Martin E.P. Seligman. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6):934, 2014.

[Peng *et al.*, 2015] Kuei-Hsiang Peng, Li-Heng Liou, Cheng-Shang Chang, and Duan-Shin Lee. Predicting Personality Traits of Chinese Users Based on Facebook Wall Posts. In *Wireless and Optical Communication Conference (WOCC), 2015 24th*, pages 9–14. IEEE, 2015.

[Pennebaker and King, 1999] James W. Pennebaker and Laura A. King. Linguistic Styles: Language Use as an Individual Difference. *Journal of Personality and Social Psychology*, 77(6):12961312, 1999.

[Pennebaker *et al.*, 2001] James W. Pennebaker, Martha E. Francis, and Roger J Booth. Linguistic Inquiry and Word Count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

[Pennebaker *et al.*, 2003] James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, 54(1):547–577, 2003.

[Pennebaker *et al.*, 2007] James W. Pennebaker, Roger J. Booth, and Martha E. Francis. Operators Manual: Linguistic Inquiry and Word Count: LIWC2007. *Austin, Texas: LIWC.*, 2007.

[Poria *et al.*, 2013a] Soujanya Poria, Alexandar Gelbukh, Basant Agarwal, Erik Cambria, and Newton Howard. Common Sense Knowledge Based Personality Recognition from Text. In *Advances in Soft Computing and Its Applications*, pages 484–496. Springer, 2013.

[Poria *et al.*, 2013b] Soujanya Poria, Alexander Gelbukh, Amir Hussain, Newton Howard, Dipankar Das, and Sivaji Bandyopadhyay. Enhanced SenticNet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, (2):31–38, 2013.

[Richardson *et al.*, 2009] John D. Richardson, John W. Lounsbury, Tripti Bhaskar, Lucy W. Gibson, and Adam W. Drost. Personality Traits and Career Satisfaction of Health Care Professionals. *The Health Care Manager*, 28(3):218–226, 2009.

[Schwartz *et al.*, 2013] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E.P. Seligman, and Lyle H. Ungar. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[Smith, 2002] Lindsay I. Smith. A Tutorial on Principal Components Analysis. *Cornell University, USA*, 51(52):65, 2002.

[Turney and Littman, 2003] Peter D Turney and Michael L Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

[Wilson, 1988] Michael Wilson. MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10, 1988.

Enhanced Sentiment Classification of Telugu Text using ML Techniques

Sandeep Sricharan Mukku
LTRC, IIIT Hyderabad
sandeep.mukku@research.iiit.ac.in

Nurendra Choudhary
LTRC, IIIT Hyderabad
nurendra.choudhary@research.iiit.ac.in

Radhika Mamidi
LTRC, IIIT Hyderabad
radhika.mamidi@iiit.ac.in

Abstract

With the growing amount of information and availability of opinion-rich resources, it is sometimes difficult for a common man to analyse what others think of. To analyse this information and to see what people in general think or feel of a product or a service is the problem of Sentiment Analysis. Sentiment analysis or Sentiment polarity labelling is an emerging field, so this needs to be accurate. In this paper, we explore various Machine Learning techniques for the classification of Telugu sentences into positive or negative polarities.

1 Introduction

Recently there is a proliferation of World Wide Web sites that emphasizes user-generated content as users are the potential content contributors. "What people think and feel" - is the important information for marketing and business operations as it makes their product or service better. Also, there are a lot of comments and blog-posts about trending activity in social media. People try to analyse this information and try to draw conclusions out of them. To better analyse and classify this information, researchers these days are actively working on sentiment analysis. Sentiment Analysis or polarity classification is an effort to classify a given text into polarities, either positive or negative. Majority of the work in the field of sentiment classification has been done in English. There has been very less contribution for regional languages, especially Indian Languages.

Telugu is a Dravidian language native to India. There are about 75 million native Telugu speakers. Telugu ranks fifteenth in the Ethnologue list of most-spoken languages worldwide[1]. Currently there are a lot of websites, blogs etc., rich in Telugu content. In our work, we tried to classify the polarity of Telugu sentences using various Machine Learning Techniques viz., Naive Bayes, Logistic Regression, SVM (Support Vector Machines), MLP (Multi Layer Perceptron) Neural Network, Decision Trees and Random Forest. We built models for two classification tasks: a binary task of classification of sentiment into positive and negative polarities and a ternary task of classification of sentiment into positive, negative and neutral polarities. The algorithm and formulation are explained in detail in later sections.

The rest of the paper is organised as follows. In section 2, we discuss the previous works and related work. In section 3, we describe the datasets used for our work. In section 4, we discuss about the methodology used in our paper which includes pre-processing, training and output. In section 5, we present the framework of our work which includes the tools and different Machine Learning techniques used in our work. In section 6, we present our experiments and discuss the results. Later, we conclude and discuss the future directions of this work.

2 Related Work

Sentiment classification is a difficult task and a lot of research has been done in the past. In this section we survey some of the methodologies and approaches used to address the task of sentiment analysis and polarity classification. Our work is motivated by most of these works.

Enhanced Naive Bayes model is used for sentiment classification task in English [Narayanan *et al.*, 2013]. Their approach is a combination of methodologies like effective negation handling, feature-selection by mutual information and word n-grams. This resulted in significant improvement of accuracy.

Learning word vectors for sentiment analysis is a research work, where Logistic Regression classifier is used as a predictor. [Maas *et al.*, 2011] proposed a methodology which can grasp both continuous and multi-class sentiment information as well as non-sentiment annotations.

[Mullen and Collier, 2004] uses support vector machines (SVMs) to bring together diverse sources of potentially pertinent information, including several favorability measures for phrases and adjectives and, where available, knowledge of the topic of the text. Predicting the helpfulness of online reviews is another area where [Lee and Choeh, 2014] uses a back-propagation multilayer perceptron neural network. This work motivated us to use multilayer perceptron (MLP) neural network for the task of sentiment classification.

Distributed Representations of Sentences and Documents is the work by [Le and Mikolov, 2014] where they make fixed length paragraph vectors or sentence vectors which are

[1] http://www.ethnologue.com/statistics/size

Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016, pages 29-34, New York City, USA, July 10, 2016.

quite useful for our work. We used the tool Doc2Vec for pre-processing the data. Further usage is explained in detail in later sections of the paper.

[Das and Bandyopadhyay, 2010] propose several computational techniques to generate sentiment lexicons in Indian languages (which includes Bengali, Hindi and Telugu languages) automatically and semi-automatically. [Das and Bandyopadhyay, 2011] proposes a tool Dr Sentiment where it automatically creates the PsychoSentiWordNet involving internet population. The PsychoSentiWordNet is an extension of SentiWordNet that presently holds human psychological knowledge on a few aspects along with sentiment knowledge.

3 Dataset

In this section, we describe the raw corpus and annotated data which are domain independent. These have been used in our experiments.

3.1 Raw Corpus

A corpus consisting of 7,21,785 raw Telugu sentences was provided by Indian Languages Corpora Initiative (ILCI)[2]. These sentences were used for training the Doc2vec model (as described in the next section) for generating sentence vectors.

3.2 Annotated Data

The corpus consists of Telugu sentences each attached with a corresponding polarity tag. There are about 1644 sentences which consists of 1068 positive, 219 negative and 357 neutral sentences. These sentences are used to train, test and evaluate the classifier models.

The corpus is prepared from raw data taken from the Telugu Newspapers[3]. This newspaper raw data was first annotated by two native Telugu speakers separately. The data was then merged by a third native speaker who also validated it simultaneously. The annotation consists of three polarity tags i.e; Positive, Negative and Neutral.

We performed inter-annotator agreement using Cohenś kappa coefficient[4]. We got the annotation consistency (k value) to be 0.92 (which is in perfect agreement).

4 Methodology

In this section we explain the steps involved in our approach. Doc2Vec tool (*Refer section 5.1*)gives the semantic representation of a sentence with respect to a dataset. This means that the vector of the sentence represents the meaning of the sentence. Therefore, classifying the semantic space according to training data can classify all the future instances of the same kind thus giving the solution to the problem of sentiment analysis.

4.1 Pre-processing

We converted the annotated data of sentences to 200-dimension feature sentence vectors. For this we used the Doc2vec tool provided by Gensim[5], a python module.

[2]http://sanskrit.jnu.ac.in/ilci/index.jsp

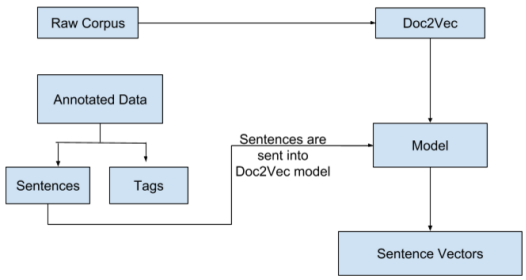
[3]http://www.w3newspapers.com/india/telugu/

[4]http://en.wikipedia.org/wiki/Cohen%27s_kappa

[5]https://radimrehurek.com/gensim/index.html

Doc2vec takes a raw corpus as input and gives us a distributional semantic representation of sentences accordingly. A Doc2vec model is trained on the raw corpus (*Refer section 3.1*). The sentences alone are taken from annotated data and passed through the trained Doc2Vec model. The model then returns sentence vectors for each of the sentences. Here we maintained the correspondence while converting between sentences and their tags.

Figure 1: Data Pre-processing

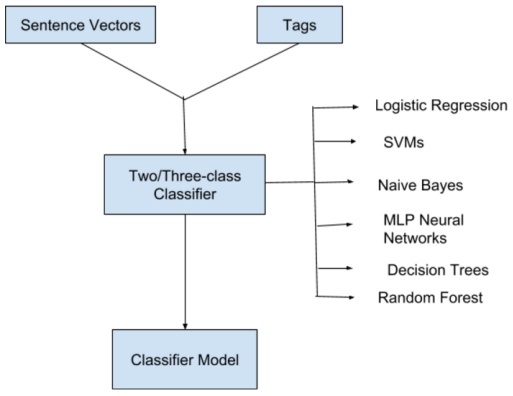


4.2 Training

In the pre-processing phase we converted each sentence of the annotated data into a sentence vector. Therefore we have a sentence vector with a corresponding tag attached to it. Hence the task is reduced to a binary or ternary classification problem. For this task we use various Machine Learning classifiers. The algorithms are explained in the following section.

The model for the classifiers are trained using sentence vectors and their corresponding tags. The models are evaluated using 5-fold cross validation where we divided the data into training and testing sets in the ratio 4:1. The model thus obtained is now ready to classify any sentence vector.

Figure 2: Training

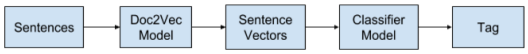


4.3 Output

In this section we discuss the final pipeline which gives the resultant tag for a given input Telugu sentence. The given input

sentence is converted into a sentence vector using a Doc2Vec model. This sentence vector is given to the trained classifier model which returns the output tag.

Figure 3: Output



5 Framework

In this section, we explain the tool used and the various Machine Learning Techniques employed.

5.1 Doc2Vec Tool

Sentence Vector is an unsupervised algorithm that learns fixed-length feature representations from variable-length pieces of texts, such as sentences. In the paper [Le and Mikolov, 2014], their algorithm represents each document by a dense vector which is trained to predict words in the document. Machine learning algorithms typically require the text to be represented as a fixed vector. Usually the most common fixed-length vector representation for texts is bag-of-words (BOW) or bag-of-n-grams [Harris, 1954]. These representations are used because they are simple and accurate. We are not using bag-of-words because this technique has many disadvantages. The word order is lost, and thus different sentences with the same set of words will have exactly the same representation. Also, we did not use bag-of-n-grams because bag-of-n-grams considers the word order in shorter context but it suffers from the curse of higher dimensionality and data sparsity. We found many advantages of sentence vectors such as learning from unlabeled data. Sentence vectors also take into consideration the word order. Doc2Vec is a tool in which sentences are converted into sentence vectors. This tool helps in pre-processing and training of data.

5.2 ML Techniques

We used scikit-learn[6] toolkit which has all these techniques pre-implemented.

Naive Bayes

Naive Bayes (NB) classifier is a probabilistic classifier which uses Bayes Theorem. This classifier evaluates the probability of an event given the probability of another event which has previously occurred. Naives Bayes classifier works very effectively for linearly separable problems. It also works fine for non-linearly separable problems.

Logistic Regression

Logistic Regression (LR) is a multi-class logistic model which is used to estimate the probability of a response based predictor variables in which there are one or more independent variables that determine an outcome. The expected values of the response based predictor variable are formed based on combination of values taken by the predictors. We took the C value (i.e. the regularization parameter) as 1.0.

[6]http://scikit-learn.org/stable/

Support Vector Machine (SVM)

SVM classifier is a supervised learning model which constructs a set of hyperplanes in a high-dimensional space which separates the data into classes. SVM is a non probabilistic linear classifier. SVM models are closely related to a Neural Network. SVM takes the input data and for each input data row it predicts the class to which this input row belongs.

Multi-Layer Perceptron (MLP) Neural Network

A multilayer perceptron (MLP) is a feed-forward artificial neural network model which maps input data sets on an appropriate set of outputs. MLP consists of multiple layers of nodes in a directed graph , each layer is fully connected to the next layer. Feed-forward means the data flows only in one direction, in our case from input to output i.e., in forward direction.

Decision Trees

Decision tree (DT) is a decision support tool that uses a tree-like model for the decisions and likely outcomes. A decision tree is a tree in which each internal (non-leaf) node is labeled with an input feature. Each leaf of the tree is labeled with a class. But for our work decision trees give less accurate results because of overfitting of training data. We took the tree depth as 20 for each decision tree.

Random Forest

Random Forest (RF) is an ensemble of Decision Trees. Random Forests construct multiple decision trees and take each of their scores into consideration for giving the final output. Decision Trees tend to overfit on a given data and hence they will give good results for training data but bad on testing data. Random Forests reduces overfitting as multiple decision trees are involved. We took the n_estimator parameter as 100.

Adaboost Ensemble

The core principle of Adaboost (A B) is to fit a sequence of weak learners (i.e., models that are only slightly better than random guessing, such as small decision trees) on repeatedly modified versions of the data. The predictions from all of them are then combined through a weighted majority vote (or sum) to produce the final prediction.

6 Experiments and Results

The method of 5-fold cross-validation is used. The experiments are performed four times (trails) to improve the validity of the results. In each experiment, the sentences in data are chosen randomly for the division into parts. These experiments are performed in the Training Step (*See Fig.2*).

The results are given below as tables. As can be observed for binary classification Random Forest, Logistic Regression and Support Vector Machines give good results. Random Forest Classifier is preferred because they have a more intuitive design and are easy-to-understand. And for ternary classification we can observe that Logistic regression gives good results.The experiments were conducted for four trials, each with five iterations (Itr) and the results are tabulated. We mentioned the average (Avg) of five iterations of each trial in the last column of each table for every technique.

6.1 Binary Sentiment Classification

The following are the accuracies where we considered only positive and negative polarities.

Trial-1

Table 1: Results of the Trial-1

	Itr 0	Itr 1	Itr 2	Itr 3	Itr 4	Avg
N B	78.55	75.93	80.12	75.60	78.65	77.77
L R	82.94	80.93	81.93	80.44	84.54	82.15
SVM	82.94	85.24	84.18	80.80	80.87	82.81
MLP	75.71	72.83	74.45	77.86	73.92	74.95
D T	70.80	73.30	71.79	68.95	72.44	71.45
R F	82.94	85.22	81.99	85.55	84.52	84.05

Trial-2

Table 2: Results of the Trial-2

	Itr 0	Itr 1	Itr 2	Itr 3	Itr 4	Avg
N B	78.55	80.42	79.96	78.91	79.20	79.41
L R	82.94	83.16	80.25	82.33	82.58	82.25
SVM	82.94	83.35	80.60	83.76	81.90	82.51
MLP	75.71	78.53	77.95	78.48	73.36	76.81
D T	68.47	66.30	71.26	69.81	68.56	68.88
R F	82.94	81.76	84.06	83.01	85.21	83.40

Trial-3

Table 3: Results of the Trial-3

	Itr 0	Itr 1	Itr 2	Itr 3	Itr 4	Avg
N B	78.55	76.10	79.05	80.02	80.10	78.77
L R	82.94	83.98	83.01	81.02	81.51	82.49
SVM	82.94	82.23	83.89	83.26	82.75	83.01
MLP	75.71	74.48	74.69	78.55	76.45	75.97
D T	71.05	72.55	70.27	71.68	73.84	71.88
R F	82.94	81.26	84.78	84.40	84.62	83.60

Trial-4

Table 4: Results of the Trial-4

	Itr 0	Itr 1	Itr 2	Itr 3	Itr 4	Avg
N B	78.55	76.47	75.58	76.30	78.78	77.14
L R	82.94	80.31	85.85	82.42	83.23	82.95
SVM	82.94	82.30	80.28	82.04	80.47	81.61
MLP	75.71	74.58	77.95	72.99	75.11	75.27
D T	68.99	68.28	70.57	68.21	67.12	68.64
R F	83.20	85.70	82.64	84.58	80.56	83.34

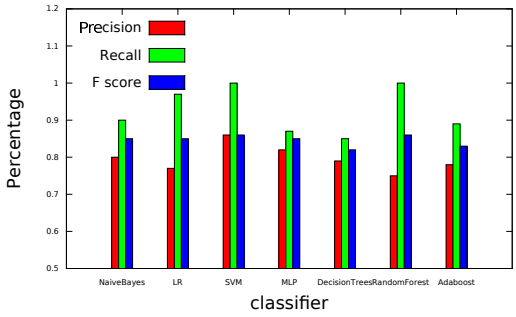
We used the ensemble of all the above six classifiers through a weighted majority vote to produce the final prediction. We get an accuracy of 73.85% for binary classification.

Precision, Recall and F-measure

Table 5: Precision, Recall and F-measure for Binary Classification

	Precision	Recall	F-measure
N B	0.80	0.90	0.85
L R	0.77	0.97	0.85
SVM	0.86	1.0	0.86
MLP	0.82	0.87	0.85
D T	0.79	0.85	0.82
R F	0.75	1.0	0.86
A B	0.78	0.89	0.83

Figure 4: Precision, Recall and F-measure for Binary classification



6.2 Ternary Sentiment Classification

There may be few cases where the data contains few sentences which may not contain any sentiment. So we considered neutral polarity leading to a ternary sentiment classification problem. The following are the accuracies where we considered all three polarities i.e., positive, negative and neutral polarities.

Trial-1

Table 6: Results of the Trial-1

	Itr 0	Itr 1	Itr 2	Itr 3	Itr 4	Avg
N B	64.54	65.1	64.85	67.42	61.96	64.77
L R	67.85	65.31	68.14	69.55	67.40	67.65
SVM	53.25	51.85	50.34	56.00	51.32	52.55
MLP	60.68	63.16	58.93	58.33	59.01	60.02
D T	52.42	54.94	50.18	54.18	52.89	52.92
R F	61.79	64.54	64.24	59.79	62.41	62.55

Trial-2

Table 7: Results of the Trial-2

	Itr 0	Itr 1	Itr 2	Itr 3	Itr 4	Avg
N B	64.54	65.66	63.85	63.59	63.37	64.20
L R	67.85	65.31	68.14	69.55	67.40	67.65
SVM	53.25	52.83	54.43	52.12	52.34	52.99
MLP	60.68	58.35	60.37	60.85	61.91	60.43
D T	54.90	51.96	55.01	52.26	53.15	53.45
R F	64.82	63.95	62.89	66.27	65.15	64.61

Trial-3

Table 8: Results of the Trial-3

	Itr 0	Itr 1	Itr 2	Itr 3	Itr 4	Avg
N B	64.54	67.17	61.80	63.78	65.54	64.56
L R	67.85	69.96	70.62	66.67	65.75	68.17
SVM	53.25	50.25	51.24	53.38	54.89	52.60
MLP	60.68	63.10	60.91	58.51	58.64	60.36
D T	51.87	50.72	50.53	51.86	51.20	51.23
R F	66.19	66.45	67.16	65.82	67.17	66.55

Trial-4

Table 9: Results of the Trial-4

	Itr 0	Itr 1	Itr 2	Itr 3	Itr 4	Avg
N B	64.54	61.86	67.20	64.18	66.48	64.85
L R	67.85	65.15	69.78	66.06	65.03	66.77
SVM	53.25	55.14	50.30	54.89	52.61	53.23
MLP	60.68	58.88	61.89	60.82	63.15	61.08
D T	51.87	52.19	53.76	54.59	52.85	53.05
R F	62.89	64.04	60.37	65.37	65.11	63.55

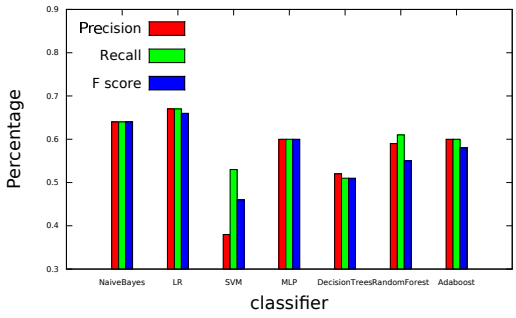
We used the ensemble of all the above six classifiers through a weighted majority vote to produce the final prediction. We get an accuracy of 60.13% for ternary classification.

Precision, Recall and F-measure

Table 10: Precision, Recall and F-measure for Ternary Classification

	Precision	Recall	F-measure
N B	0.64	0.64	0.64
L R	0.67	0.67	0.66
SVM	0.38	0.53	0.46
MLP	0.60	0.60	0.60
D T	0.52	0.51	0.51
R F	0.59	0.61	0.55
A B	0.60	0.60	0.58

Figure 5: Precision, Recall and F-measure for Ternary classification



7 Conclusion

Telugu is an agglutinative language. Considering this fact, we have achieved good results. Sentiment Analysis has not yet been tried on agglutinative Dravidian Languages. Since our work is the first attempt of this kind, we are not able to discuss comparative results. This approach produces a more focused and accurate sentiment summary of a given Telugu sentence which is useful for the users. This approach is not restricted by any domain. However, small modifications in the pre-processing would be sufficient to use this algorithmic formulation in different domains or languages.

Future Work

- To build a dictionary of frequently occurring positive and negative words and construct a lexicon-based system using it.
- To integrate a Morph Analyser to address the issue of agglutination.
- To test the system for different Indian languages.
- To work on the trending code-mixed data.

References

[Bakliwal *et al.*, 2011] Akshat Bakliwal, Piyush Arora, Ankit Patil, and V Verma. Towards enhanced opinion classification using nlp techniques. In *Proceedings of the 5th international joint conference on natural language processing (IJCNLP). Chiang Mai, Thailand*, pages 101–107. Citeseer, 2011.

[Balamurali, 2012] AR Balamurali. Cross-lingual sentiment analysis for indian languages using linked wordnets. 2012.

[Das and Bandyopadhyay, 2010] Amitava Das and Sivaji Bandyopadhyay. Sentiwordnet for indian languages. *Asian Federation for Natural Language Processing, China*, pages 56–63, 2010.

[Das and Bandyopadhyay, 2011] Amitava Das and Sivaji Bandyopadhyay. Dr sentiment knows everything! In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies: systems demonstrations*, pages 50–55. Association for Computational Linguistics, 2011.

[Das *et al.*,] Dipankar Das, Soujanya Poria, Chandra Mohan Dasari, and Sivaji Bandyopadhyay. Building resources for multilingual affect analysis–a case study on hindi, bengali and telugu. In *Workshop Programme*, page 54. Citeseer.

[Go *et al.*, 2009] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1:12, 2009.

[Harris, 1954] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[Joshi *et al.*, 2010] Aditya Joshi, AR Balamurali, and Pushpak Bhattacharyya. A fall-back strategy for sentiment analysis in hindi: a case study. *Proceedings of the 8th ICON*, 2010.

[Le and Mikolov, 2014] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

[Lee and Choeh, 2014] Sangjae Lee and Joon Yeon Choeh. Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Systems with Applications*, 41(6):3041–3046, 2014.

[Liu, 2010] Bing Liu. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666, 2010.

[Liu, 2012] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[Maas *et al.*, 2011] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 142–150. Association for Computational Linguistics, 2011.

[Mullen and Collier, 2004] Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources. In *EMNLP*, volume 4, pages 412–418, 2004.

[Narayanan *et al.*, 2013] Vivek Narayanan, Ishan Arora, and Arjun Bhatia. Fast and accurate sentiment classification using an enhanced naive bayes model. In *Intelligent Data Engineering and Automated Learning–IDEAL 2013*, pages 194–201. Springer, 2013.

[Pang and Lee, 2008] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[Patra *et al.*, 2015] Braja Gopal Patra, Dipankar Das, Amitava Das, and Rajendra Prasath. Shared task on sentiment analysis in indian languages (sail) tweets-an overview. In *Mining Intelligence and Knowledge Exploration*, pages 650–655. Springer, 2015.

[Wilson *et al.*, 2005] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

A Hybrid Approach based Sentiment Extraction from Medical Contexts

[1]Anupam Mondal [2]Ranjan Satapathy [1]Dipankar Das [1]Sivaji Bandyopadhyay
[1]Computer Science and Engineering, Jadavpur University, India
[1]anupam@sentic.net, [1]ddas@cse.jdvu.ac.in, [1]sbandyopadhyay@cse.jdvu.ac.in
[2]School of Computer and Information Sciences, University of Hyderabad, India
[2]kumarsatpathy@gmail.com

Abstract

In the domain of Bio medical Natural Language Processing (Bio-NLP), the information extraction and context sentiment identification are treated as emerging tasks. Several linguistic features like negation, uni-gram, bi-gram, Part-of-Speech (POS) have been used to extract the medical concepts and their sense-based context level information. Thus, in the present attempt, a hybrid approach which is the combination of both linguistic and machine learning approaches has been introduced to extract the contextual sense-based information from a medical corpus. The extraction of sentiment oriented keywords is the crucial part towards identifying the senses of medical contexts. In our previous work, we have developed a medical sense-based lexicon known as WordNet of Medical Event (WME). Several sentiment lexicons like SentiWordNet, SenticNet etc. were used to represent WME. In contrast, one of our primary motivations here is to build a sentiment extraction model based on medical contexts to leverage the knowledge of WME using a hybrid approach. The developed model is based on two phases, namely preprocessing phase and learning phase. The preprocessing phase is responsible for extracting and preparing structural data from the raw contexts whereas the learning phase helps to identify the sentiment patterns and evaluate the sentiment extraction process. The two phased hybrid model provides us 81% accuracy for extracting the sentiment based medical contexts as positive and negative by employing NaïveBayes and Sequential minimal optimization (SMO) supervised classifiers.

1 Introduction

One of the major objectives of Sentiment Analysis is to identify and extract the subjective information from a given text using rule based or machine learning approaches [Cambria, 2016]. The domain specific knowledge with above mentioned approaches help us to extract the contextual sentiment information from the medical corpus. Due to lack of involvement of domain experts and unavailability of domain specific structured corpus, the task is challenging in Bio-NLP domain. To overcome the scarcity of such domain specific knowledge for sentiment analysis, several lexicons have been developed like Medical Event Net (MEN), Medical Fact Net (MFN), Medical Belief Net (MBN) and WordNet of Medical Event (WME) [Cambria *et al*., 2010]. These lexicons help to extract the sense of a medical concept, fact and belief oriented information. The present paper reports the development of a medical context based sentiment extraction model. Hence, one of our primary aims is to identify the sense-based concepts from the medical contexts and extract their related sentiment features. In order to identify the sense-based medical concepts, we have introduced the current version of WordNet of Medical Event (WME2.0) knowledge base. WME2.0 contains the medical concept information with their related linguistic and sense-oriented features like POS, gloss of the concept, semantics, polarity score, affinity score, gravity score and sense(s). Among all these features, we have only considered the sense-based features like semantics, polarity score, affinity score and sense to develop our present sentiment extraction model [Swaminathan *et al*., 2010]. On the top of extracted medical concepts based on WME2.0 lexicon, we have applied linguistic and machine learning approaches to get the final sentiment of the contexts. The linguistic approach helps to manage the negation of the contexts as well as derive new rules to extract the sense(s) of such contexts. The POS, unigram, bi-gram, affinity score, polarity score and sense features of the medical concepts of WME2.0 help to extract the sentiment of the medical contexts. The supervised machine learning approach has been introduced to verify the contextual sentiment extracted using linguistic approach. In the process, we have applied NaïveBayes and Sequential minimal optimization (SMO) supervised machine learning classifiers on the derived linguistic features.

In the paper, we have incorporated both linguistic and machine learning approaches together as a hybrid model to leverage the sentiment oriented knowledge of both the domain [Villena-Romn *et al*., 2011]. The proposed hybrid model follows two phase architecture namely preprocessing phase and learning phase. In pre-processing phase, we have focused on the preparation of structured medical concepts from the raw medical contexts and the

Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016, pages 35-40, New York City, USA, July 10, 2016.

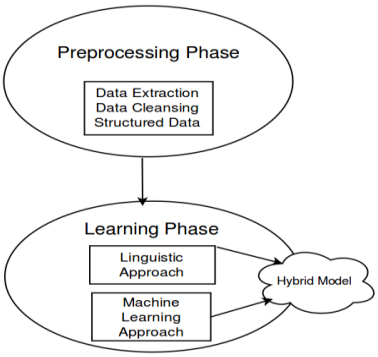
learning phase helps to extract the sentiment of such contexts and evaluate them. The two phase model generates the output in the form of positive or negative sentiment of the context. The hybrid approach based learning phase provides 81% accuracy to extract the medical context based sentiment information.

The remainder of the paper is structured as follows, Section 2 presents related work followed by model design describing the pre-processing and learning phases in Section 3. Section 4 talks about the model discussion and evaluation process we have followed in the paper. Finally, in Section 5, we present our conclusion and future scopes of the model.

2 Related Work

Sentiment analysis of medical contexts is contributory and growing research field under Bio-NLP domain [Cambria *et al*., 2013]. A large number of unstructured corpora and lack of domain experts' involvement have introduced more challenge in this task. In the process, the researchers focused on developing medical sentiment-based lexicon to identify the sentiments of medical concepts. Therefore, the medical concepts and their sense based features indeed help to identify the sentiment of the medical contexts. The linguistic, machine learning and hybrid approaches have been introduced to build the concept and context based sentiment extraction systems. The linguistic approach helps to find the negation words, phrases and construct the knowledge-based rules (with unigram, bigram and n-gram features) for the context level sentiment extraction [Elkin *et al*., 2005; Niu *et al*., 2005; Szarvas *et al*., 2008]. Smith and Fellbaum, 2004 developed a Medical Word-Net (MEN) along with two subnetworks, namely Medical FactNet (MFN) and Medical BeliefNet (MBN), for the evaluation of consumer health reports [Smith and Fellbaum, 2004]. MEN was developed with the help of formal architecture of the Princeton WordNet [Fellbaum, 1998]. MFN serves to assist the non-expert group in providing a better understanding of basic medical information. MBN identifies beliefs about the medical phenomenon. Their primary motivation was to develop a network of medical information retrieval systems with visualization effect. The domain-specific knowledge and the abovementioned features are essential to improve the efficiency of the sentiment extraction system [Shukla *et al*., 2015]. So, these approaches were not able to provide adequate accuracy due to the lack of knowledge involvement from the domain experts. Hence, to overcome the mentioned problem, the researchers introduced supervised machine learning approaches [Smith and Lee, 2012]. Standard NaïveBayes, Multinomial NaïveBayes and Support Vector Machine (SVM) supervised classifiers were applied with unigram, bigram, Parts Of Speech (POS) and negation features under the machine learning framework. The researchers have also used hybrid approaches to improve the accuracy of the medical context based sentiment extraction systems. One of the hybrid approaches was developed with the combination of linguistic and machine learning approaches [Boytcheva *et al*., 2005; Villena-Romn *et al*., 2011]. Sohn *et al*., 2012, developed an emotion identification system from suicide notes using the hybrid approach [Sohn *et al*., 2012]. The suicide notes were provided by the challenge organizers of Informatics for Integrating Biology and the Bedside (I2B2). Machine learning, linguistic rule-based and their combined approaches have been applied to the training dataset of the suicide notes and the system provided 0.5640 micro-average F-score for the training dataset. Birks et al., 2009, applied the combination of RIPPER (Repeated Incremental Pruning to Produce Error Reduction), multinomial NaïveBayes classifier and manual pattern matching rules to identify the emotions of the sentences [Birks *et al*., 2009]. Mondal et al., 2016, developed WordNet of Medical Events (WME) lexicon to identify the medical concepts and their knowledge-based and semantic features using hybrid approach [Mondal *et al*., 2015]. The latest version of WME (WME2.0) contains POS, semantics, gloss, affinity score, gravity score, polarity score and sense features of the concepts [Mondal *et al*., 2016]. WME2.0 sentiment lexicon has identified the senses of the concepts using SentiWordNet[1], SenticNet[2], BingLiu[3] and Taboda's adjective list [Mondal *et al*., 2016; Mondal *et al*., 2015; Taboada *et al*., 2011]. In this paper, we have used the WME2.0 lexicon to identify the concepts and their features to extract sentiments of the medical contexts.

Figure 1: Two phase proposed Model



3 Model Design

The knowledge-based sentiment lexicon is crucial to design a context based sentiment extraction system. The medical concepts and their linguistic features are extracted from the domain-specific sentiment lexicon. To overcome the problem of experts' availability, we have formulated WME2.0 lexicon with a hybrid approach. It adds an extra dimension

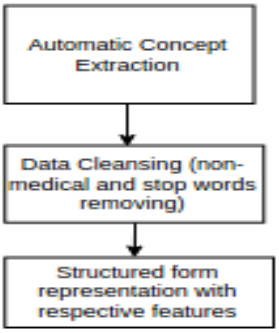
[1] http://sentiwordnet.isti.cnr.it/
[2] http://sentic.net/
[3] https://www.cs.uic.edu/liub/FBS/sentiment-analysis.html

for improving the accuracy of the extracted medical context sentiment. The proposed hybrid approach is the combination of linguistic and machine learning approach. The approach consists of two phases namely pre-processing and learning phase. Figure 1 shows the architecture of the proposed approach (model).

3.1 Pre-processing phase

The phase extracts the sentiments of medical contexts in the form of context related medical concepts, their sentiments and knowledge-based information. The structured form of the concepts is essential in identifying the important medical concepts from the context.

Figure 2: Flowchart of Preprocessing Phase



In this concern, to represent the structured medical concepts, the required steps are data extraction, cleansing and formatting. The research community provided various linguistic resources such as open source data preprocessing tools (viz. NLTK, stemming etc.) [Na *et al*., 2012]. The following steps illustrate the basic operations of the pre-processing phase:

Data Extraction: The medical concepts extraction from a given context is the primary task of this step. WME2.0 helps to extract the medical concepts and their linguistic and sense-based features from the context. Moreover, the non-medical concepts and their sense identification are also essential to identify the sentiment of the context. The non-medical concepts the senses have been extracted using SentiWordNet and SenticNet lexicons [Cambria *et al*., 2014; Cambria *et al*., 2013; Esuli and Sebastiani, 2006].

Data Cleansing: Data cleansing step is responsible to remove the context related stop-words and stemmed the concept words. The classification of medical and nonmedical concepts and identification of negation words (like no, not, never etc.) are also taken care of by data cleansing step [Huang and Lowe, 2007].

Data Formatting: Data formatting has been applied to represent the structured form of the extracted medical concepts [Hussain *et al*., 2011]. The extracted structured (vector) concepts have been forwarded to the learning phase along with their features. The concept structure is represented as follows:

<Concept (gastric), POS (noun), Semantics (abdominal breathing, visceral, intestinal, belly, duodenal, stomachic), Polarity Score (-0.5), Sense (Negative)>

3.2 Learning phase

Followed by the pre-processing phase, the hybrid approach has been introduced in the learning phase to build the contextual sentiment extraction system. Linguistic and machine learning has been combined to form the hybrid approach. The linguistic approach with WME2.0 knowledge base lexicon helps to identify the hidden rules. These rules are able to extract the concept sentiment and their polarity. The extracted linguistic concept features (rules) were fed to the supervised machine learning classifiers to evaluate the accuracy of the model. The linguistic approach provides a support to handle the negation effect of the context and help to identify the appropriate sentiment of the context [Huang and Lowe, 2007]. The learning phase is illustrated as follows:

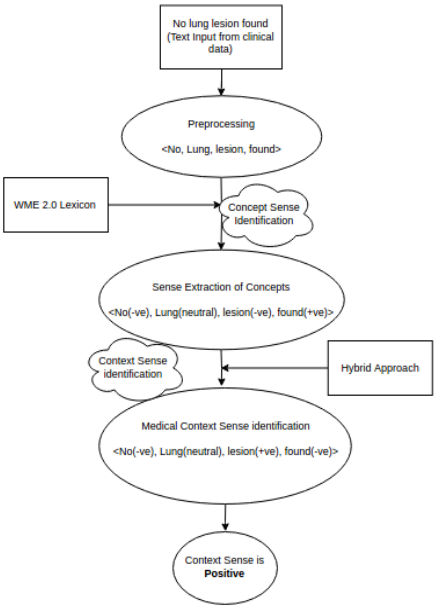
Step 1: Identify the polarity score and sense of each concept (medical and non-medical) of the context.
Step 2: Linguistic approach-based negation words (concept) handling.
Step 3: Calculate the overall polarity of the context.
$\text{Context}_{\text{polarity}} = \sum \text{Polarity}_c$
Where, c = number of concepts in the context and Polarity_c indicates the polarity score of each concept.
Step 4: The context sentiment has been evaluated using $\text{Context}_{\text{polarity}}$ score.

4 Discussion and Evaluation

The context related medical concepts and their semantic features (extraction polarity, semantics and sense) are required to identify the sentiment of the medical context [Sarker *et al*., 2011]. In the process, the statistical and linguistic features based medical sentiment lexicons were facing difficulties due to the unstructured nature of the corpus. So, the researchers tried to build an intelligent automated sentiment extraction system in the Bio-NLP domain [Shukla *et al*., 2015; Sohn *et al*., 2012]. The system helps to extract the structured knowledge-based information with a proper sentiment of the context. WordNet of Medical Event (WME2.0) was introduced to identify the medical concept and their sense-based features. The WME2.0 lexicon able to extract the medical concepts and their POS, semantics, gloss, affinity score, gravity score, polarity score and sense. On the top of WME2.0 lexicon, the hybrid approach has been applied to extract the context level sentiment for the

proposed model. The model is based on two phases namely pre-processing and learning phase. The pre-processing phase has considered the concept extraction (medical and non-medical concept), concept cleansing (concept stemming and stop-words removing) and concept formatting <Concept, POS, semantic, polarity score, sense> steps. The learning phase identified the sentiment using the linguistic and machine learning approaches on the pre-processing step driven data. The concept linguistic features and knowledge based WME sentiment resource help to extract the overall context sentiment and polarity score. The linguistic approach provides a support to handle the negation and identifies the correct sense of the context. The medical context "No lung lesion found" has been evaluated as "positive" sentiment after handling the negation. The system first extracts the concepts and their sense as "no (-ve)", "lung (neutral)", "lesion (-ve)" and "found (+ve)" using WME2.0 resource. The linguistic-based negation handling approach has been applied on the extracted sense and identify the overall context sense as "positive". In the learning phase, the hybrid approach has been introduced to extract and measure the accuracy of the context sentiment. The linguistic approach involves knowledge-based medical concept mapping with WME2.0 lexicon. Further, the NaïveBayes and Sequential minimal optimization (SMO) support vector based supervised machine learning approaches have been employed for evaluating the accuracy of the model. Figure 3 and Figure 4 describe the positive and negative contexts with respect to the sentiment extraction process, respectively.

Figure 3: Positive Sentiment extraction



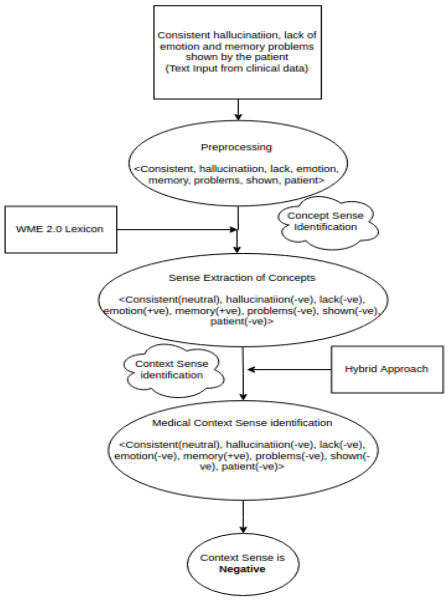
4.1 Evaluation Process

To develop and measure the accuracy of the context level sentiment extraction system, the data has been collected from the open source resource[4]. We have extracted 7042 number of medical contexts and applied through the proposed sentiment extraction system. The context sentiment extraction system has provided 3265 number of the positive and 3777 number of the negative sentiments of the contexts. To evaluate the extracted context sentiment, the linguistic features (number of negation word, context polarity score and sense) were fed to the NaïveBayes and support vector based SMO supervised machine learning classifiers under the WEKA[5] tool. The extracted 7042 number of context data has been represented as 4900 number of training and the remaining 2142 number of test dataset. The system's accuracy was measured as F-Measure with four types of models like, Use training set, Supplied test set, Cross-validation Folds 10 and Percentage split %66. Table 1 shows the F-Measures of these modes for the NaïveBayes and support vector based SMO supervised classifiers. The linguistic and machine learning based hybrid approach provides the accuracy score nearly 81% for the medical context sentiment extraction model.

Table 1: F-Measure of Supervised classifiers

Model	NaïveBayes	SMO
Use training set	0.868	0.890
Supplied test set	0.815	0.815
Cross-validation Folds 10	0.864	0.867
Percentage split %66	0.873	0.879

Figure 4: Negative Sentiment extraction



[5] http://weka.wikispaces.com/

5 Conclusion and Future scope

Sentiment or opinion analysis is important to extract the contextual information from the medical context under NLP domain. The context sentiment helps to identify the knowledge based information and proper utilization of the context. The paper has reported a hybrid approach based context sentiment extraction model with two phases. The phases are preprocessing (important medical keywords extraction) and learning (respective sentiment identification). In the process, the linguistic and machine learning combined hybrid approach has been applied on the top of WordNet of Medical Event (WME2.0) lexicon to extract the medical concepts in order to identify the sentiment of the medical context. The medical concept polarity score and their related sense helps to identify the medical context sentiment [Cambria, 2013] and [Cambria *et al*., 2015]. WME2.0 lexicon driven medical concepts affinity score and their semantic features are crucial in building the proposed model. The medical concept semantics, polarity score and affinity score helps to identify the medical concept sentiment with polarity score. The hybrid approach provides nearly 81% accuracy for the proposed context sentiment extraction system. Hence, the future research will focus to develop some practical applications relating to the current work as medical annotation and context summarization system. These systems will provide the support to the expert and non-expert groups in their respective applications.

References

[Mondal *et al*., 2016] Anupam Mondal, Dipankar Das, Erik Cambria and Sivaji Bandyopadhyay. WME: Sense, polarity and affinity based concept resource for medical events. In *Proceedings of the Eighth Global WordNet Conference*, pages 242–246, 2016.

[Birks *et al*., 2009] Yvonne Birks, Jean McKendree, and Ian Watt. Emotional intelligence and perceived stress in healthcare students: a multi-institutional, multi-professional survey. *BMC Medical Education*, 9(1):1–8, 2009.

[Boytcheva *et al*., 2005] Svetla Boytcheva, Albena Strupchanska, Elena Paskaleva, Dimitar Tcharaktchiev, and Dame Gruev Str. Some aspects of negation processing in electronic health records. In *Proceedings of International Workshop Language and Speech Infrastructure for Information Access in the Balkan Countries*. Pages 1—8, 2005.

[Cambria *et al*., 2010] E. Cambria, A. Hussain, T. Durrani, C. Havasi, C. Eckl, and J. Munro. Sentic computing for patient centered applications. In *IEEE 10th International Conference on Signal Processing Proceedings*, pages 1279–1282, Oct 2010.

[Cambria, 2013] Erik Cambria. An introduction to concept-level sentiment analysis. In *Advances in Soft Computing and Its Applications - 12th Mexican International Conference on Artificial Intelligence*, MICAI 2013, Mexico City, Mexico, November 24-30, 2013, Proceedings, Part II, pages 478–483, 2013.

[Cambria, 2016] Erik Cambria. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107, 2016.

[Cambria *et al*., 2015] Erik Cambria, Jie Fu, Federica Bisio, and Soujanya Poria. Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, January 25-30, 2015, Austin, Texas, USA, pages 508–514, 2015.

[Cambria *et al*., 2014] Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. Senticnet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In *AAAI Conference on Artificial Intelligence*, 2014.

[Cambria *et al*., 2013] Erik Cambria, Bjrn Schuller, Yunqing Xia, and Catherine Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21, 2013.

[Hussain *et al*., 2011] Hussain A Cambria E and Eckl C. Bridging the gap between structured and unstructured health- care data through semantics and sentics. In *Proceedings of ACM WebSci*, Koblenz, 2011.

[Elkin *et al*., 2005] Peter L. Elkin, Steven H. Brown, Brent A. Bauer, Casey S. Husser, William Carruth, Larry R. Bergstrom and Dietlind L. Wahner-Roedler. A controlled trial of automated classification of negation from clinical notes. *BMC Medical Informatics and Decision Making*, 5(1):1–7, 2005.

[Esuli and Sebastiani, 2006] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06)*, pages 417–422, 2006.

[Fellbaum, 1998] Christiane Fellbaum. WordNet: an electronic lexical database. *MIT Press*, 1998.

[Huang and Lowe, 2007] Yang Huang and Henry J. Lowe. A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association: JAMIA*, 14(3):304–311, May 2007.

[Mondal *et al*., 2015] Anupam Mondal, Iti Chaturvedi, Dipankar Das, Rajiv Bajpai, and Sivaji Bandyopadhyay. Lexical resource for medical events: A polarity based approach. In *IEEE ICDM Workshops*, pages 1302–1309. IEEE, 2015.

[Na *et al*., 2012] Jin-Cheon Na, Wai Yan Min Kyaing, Christopher SG Khoo, Schubert Foo, Yun-Ke Chang, and Yin-Leng Theng. Sentiment classification of drug reviews using a rule-based linguistic approach. In *The outreach of digital libraries: a globalized resource network*, pages 189–198. Springer, 2012.

[Niu *et al*., 2005] Yun Niu, Xiaodan Zhu, Jianhua Li, and Graeme Hirst. Analysis of polarity information in medical text. In *Proceedings of the American Medical Informatics Association Annual Symposium*, 2005.

[Sarker *et al*., 2011] Abeed Sarker, Diego Moll´a-Aliod, C´ecile Paris, et al. Outcome polarity identification of medical papers. *Melbourne: Australian Language Technology Association*. 2011.

[Shukla *et al*., 2015] Ravi Shankar Shukla, Kamendra Singh Yadav, Syed Tarif Abbas Rizvi, and Faisal Haseen. An Efficient Mining of Biomedical Data from Hypertext Documents via NLP. In *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014*: Volume 1, pages 651–658. Springer International Publishing, Cham, 2015.

[Smith and Fellbaum, 2004] Barry Smith and Christiane Fellbaum. Medical wordnet: A new methodology for the construction and validation of information resources for consumer health. In *Proceedings of COLING*, 2004.

[Smith and Lee, 2012] Phillip Smith and Mark Lee. Cross-discourse development of supervised sentiment analysis in the clinical domain. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, WASSA '12*, Association for Computational Linguistics, pages 79–83, Stroudsburg, PA, USA, 2012.

[Sohn *et al*., 2012] Sunghwan Sohn, Manabu Torii, Dingcheng Li, Stephen Wu, Hongfang Liu, and Avishwar Wagholikar. A Hybrid Approach to Sentiment Sentence Classification in Suicide Notes. In *Biomedical Informatics Insights*, pages 43+, January 2012.

[Swaminathan *et al*., 2010] Rajesh Swaminathan, Abhishek Sharma, and Hui Yang. Opinion mining for biomedical text data: Feature space design and feature selection. In *The Nineth International Workshop on Data Mining in Bioinformatics, BIOKDD*, 2010.

[Szarvas *et al*., 2008] Gy¨orgy Szarvas, Veronika Vincze, Rich´ard Farkas, and J´anos Csirik. The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, Association for Computational Linguistics, pages 38–45, Columbus, Ohio, June 2008.

[Taboada *et al*., 2011] Maite Taboada, Milan Tofiloski, Julian Brooke, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. Journal of *Computational linguistics*, volume 37, number 2, pages 267-307, publisher MIT Press, 2011.

[Villena-Romn *et al*., 2011] Julio Villena-Romn, Sonia Collada-Prez, Sara Lana-Serrano, and Jos Carlos Gonzlez Cristbal. Hybrid approach combining machine learning and a rule-based expert system for text categorization. In *R. Charles Murray and Philip M. McCarthy, editors, FLAIRS Conference*. AAAI Press, 2011.

Praiseworthy Act Recognition Using Web-based Knowledge and Semantic Categories

Rafal Rzepka, Kohei Matsumoto and Kenji Araki
Graduate School of Information Science and Technology
Hokkaido University, Japan
{rzepka,matsumoto,araki}@ist.hokudai.ac.jp

Abstract

In this paper we[1] introduce our novel method for utilizing web mining and semantic categories for determining automatically if a given act is worth praising or not. We report how existing lexicons used in affective analysis and ethical judgement can be combined for generating useful queries for knowledge retrieval from a 5.5 billion word blog corpus. We also present how semantic categorization helped the proposed method to finally achieve 94% of agreement with human subjects who decided which act, behavior or state should be praised. We also discuss how our preliminary findings might lead to developing an important social skill of a robotic companion or an automatic therapist during their daily interaction with children, elderly or depressed users.

1 Introduction

Predictions from world demographic trends show that the current ratio of people aged sixty or more (12.6%) will nearly double in 2050 (almost 22%)[2]. Younger generations would need to work more and worry more, not only about their aged parents but also about their children to whom they would dedicate less time. Stress among working age group could be caused not only by work itself but also by the awareness of children and parents often left to their own devices. Data gathered by American Depression and Bipolar Support Alliance[3] indicates that depression most often strikes at age 32 in the United States, but poses also an obvious problem among different age groups. One child in 33 children and one in eight adolescents have clinical depression and even if as many as six million elderly people are affected by mood disorders, but only 10% ever receive treatment. Precise numbers are often difficult to obtain as many subjects do not want to participate in studies, do not respond to surveys, do not answer the door or have insufficient language abilities[4]. Problems related to psychological disorders could be alleviated by technological advancements, including progress in Artificial Intelligence, especially in cases of social withdrawal in which depressed adolescents prefer to deal with computers than with people. As psychology studies show [Hofmann *et al.*, 2012], the depression can be treated by cognitive behavioral therapies (CBT) as efficiently as medicaments and such treatment is based on conversation. Although computers are already used as supportive tools in CBT [Wright *et al.*, 2005], we are far away from entrusting patients to autonomous therapists. However, we believe that various conversational rules utilized in dialog-based therapies and other positive aspects [Burnard, 2003; Zimmerman *et al.*, 2009] of a conversation itself can be implemented in artificial agents like companion robots [Sarma *et al.*, 2014]. In this paper we introduce our idea how to utilize Natural Language Processing techniques, a set of lexicons and semantic categories to web mine knowledge necessary for recognizing if an action being a dialog topic should be e.g. complimented by an agent.

1.1 Importance of Praising

We chose the act of praising to be implemented in our artificial agent for a variety of reasons. First of all it is an evaluation task which positively influences a praised person [Kanouse *et al.*, 1981] and motivates, especially children [Henderlong and Lepper, 2002]. Often seen in interpersonal interaction, praising is used to encourage others, to socialize, to integrate groups, and to influence people [Lipnevich and Smith, 2008]. It is believed to have beneficial effects on self-esteem, motivation and performance [Weiner *et al.*, 1972; Bandura, 1977; Koestner *et al.*, 1987]. It is widely acknowledged that to praise oneself could substantially help dealing with depression [Swann *et al.*, 1992] and praising improves behavior [Garland *et al.*, 2008], academic performance [Strain *et al.*, 1983] and work performance [Crowell *et al.*, 1988]. But there is some other interesting and difficult aspect of praising – the praiser has to be competent and share some relationships with the praised person [Carton, 1996]. Also, from the Artificial Intelligence point of view, the automatic distinction between praiseworthy and not praiseworthy

[1] Second author is currently with Panasonic Co.
[2] www.unfpa.org/ageing
[3] www.dbsalliance.org
[4] www.nimh.nih.gov/health/statistics/prevalence/major-depression-among-adults.shtml

Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016, pages 41-47, New York City, USA, July 10, 2016.

acts is an interesting long-term challenge to create a righteous and trustful machine and, in this particular case, to investigate if the Web resources could become a sufficient knowledge base for such tasks. Our hypothesis is that knowing the polarity of consequences of human acts might be the key to an automatic evaluation of these acts.

1.2 State of the Art

The authors have found only one research proposal dedicated specifically to automating praising. In 1998 [Tejima *et al.*, 1998] have published a two page paper in which they describe their observations from physiotherapists' sessions with elderly. The researchers proposed a simple verbal encouragement algorithm for walking training and implemented it later [Tejima and Bunki, 2001], however the effectiveness could not be confirmed due to the insufficient number of experimental subjects. Causing positive moods in interlocutors can be found as a sub-task in Human-Computer Interaction (HCI) field, especially in learning-oriented agents [Fogg and Nass, 1997; Kaptein *et al.*, 2010] but the studies utilize scenarios and manually created rules when to praise. Systems that accept, in theory, any sentence as an input and recognize polarity or emotive categories were proposed in the fields of sentiment analysis and affect recognition [Wilson *et al.*, 2005; Strapparava and Mihalcea, 2008] and the basic idea for our system is borrowed from their approaches. However these methods cannot be utilized straightforwardly because *being positive* does not have to mean an act is *worth praising* ("I saw a movie" is labelled positive by these methods but it usually does not mean we need to react with a compliment to such a statement). For English language there are promising methods for retrieving *goodFor* and *badFor* events [Deng and Wiebe, 2014] and for acquiring knowledge of stereotypically positive and negative events from personal blogs [Ding and Riloff, 2016]. Basically any new trend in the field [Cambria *et al.*, 2013; Socher *et al.*, 2013] should eventually help improve our results as soon as they are implemented for Japanese language, which often has much less resources to keep up with the latest methods. For Japanese [Rzepka and Araki, 2015] have proposed a system that evaluates textual inputs from a moral perspective. Similarly to our approach it uses lexicons and one of them, based on Kohlberg's theory of moral stages development [Kohlberg, 1981], includes praise-punishment polarized pairs. However, the lexicon contains only 14 praise related words limited to synonyms of the verb "praise" which, as shown later in the comparison experiment, are insufficient for our purposes.

2 System Overview

The algorithm of our system is presented in Figure 1. In the first step an input act (noun - verb pair we treat as the minimal semantic unit describing any act) in Japanese language is morphologically analyzed by MeCab[5] to determine a noun, a verb and the joining particle representing grammatical case (e.g. *aisatsu-o suru* "to greet someone" or *yakusoku-o mamoranai* "not keeping promises", where particle "o" indicates an object of given verb). Then the system adds to the verb 15 suffixes representing conditions and temporal sequences to retrieve more adequate sentences (*waruguchi itta ato* "after calling names", *waruguchi iu toki* "when called names", *waruguchi itte* "called names and then", etc.). Because particles are often omitted in colloquial Japanese, another set of 15 phrases without particles is created and the final 30 phrases together with phrases with verbs in their basic (dictionary) form become queries for 5.5 billion word YACIS corpus of Japanese raw blogs [Ptaszynski *et al.*, 2012]. Text retrieved from the corpus is then cleaned – emoticons usually used as sentence boundaries are converted to fullstops and too long and too short sentence candidates are deleted. In the next step, the generated temporary corpus of sentences containing input acts is normalized to verb dictionary forms and divided into meaningful chunks by Argument Structure Annotator ASA [Takeuchi *et al.*, 2010] to avoid granular division of morphological analyzer. For instance "was | beat | ing | brother" becomes "beat brother" and such transitions are made to increase the coverage of matching chunks with phrases from lexicons in the next step. Every match is scored 1 and the totals are compared. If there are more than 50% of positive or negative counts, the act is estimated as praiseworthy or not praiseworthy accordingly. Although in morality estimation task 60/40 ratio scored highest [Rzepka and Araki, 2015], in our task the 50/50 ratio achieved better results.

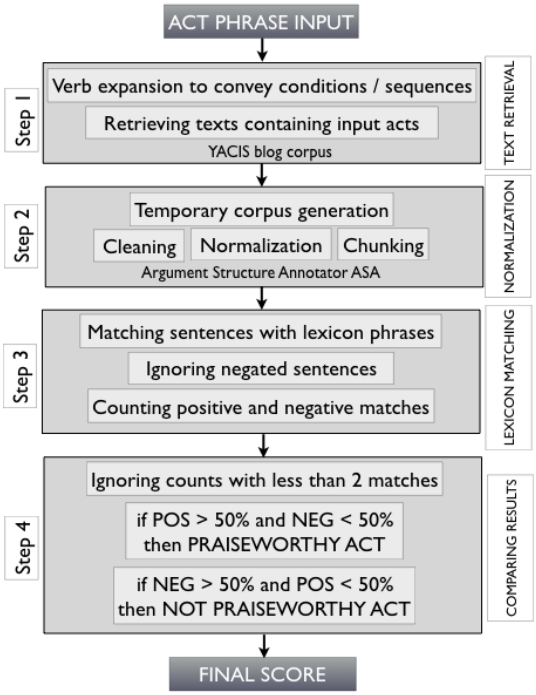


Figure 1: Algorithm for retrieving and analyzing consequences of acts in order to determine if they should be praised.

[5] taku910.github.io/mecab/

2.1 Lexicons

As mentioned in the introduction, we hypothesized that measuring the polarity of act consequences might be the key for recognizing praiseworthy acts. Although aware of possible problems mentioned in the Introduction, we decided to investigate how efficient the existing emotional recognition methods could deal with our task. Therefore firstly we chose two different freely available lexicons used for lexicon based polarity recognition in Japanese language. The larger one was statistically generated from manually annotated sentences in the study of [Takamura *et al.*, 2005]. It contains 55,102 words divided into positive (5,121 words) and negative (49,981 words) ones. Every word was automatically scored on the scale from minimal -1 to maximal 1 and the words closer to 0 tend to be inaccurately labeled (e.g. *okaasan* "mom" or *narubeku* "as possible", are marked as negative words), therefore using the whole (significantly unbalanced) set would cause drops in accuracy. In order to minimize this problem and to make the lexicon more balanced, after analyzing the entries we used most positive 3,000 and most negative 3,000 words (closest to 1 and -1 from each side) and called it "Statistical Lexicon".

Another lexicon used in polarity detection in Japanese texts is created manually by [Nakamura, 1993] from emotive sentences retrieved from Japanese literature. The words are separated into ten categories (Like, Joy, Relief, Dislike, Anger, Fear, Shame, Sadness, Excitement, Surprise) and because Excitement and Surprise have no distinct valence, these two categories were excluded. The combined words from Like, Joy and Relief form a positive subset and Dislike, Anger, Fear, Shame and Sadness form a negative one. Resulting lexicon of 526 positive and 756 negative words (1,282 in total) we call here "Literature Lexicon" to make it more comprehensible while presenting comparison between lexicons.

As mentioned before, a positive act does not necessarily imply being praiseworthy, therefore we decided also to test a lexicon used for ethical judgement by [Rzepka and Araki, 2015]. This relatively small set, containing 65 positive and 69 negative words (134 in total), was created by applying phrases related to the five stages of moral development proposed by [Kohlberg, 1981]: obedience / punishment, self-interest, social norms, authority / social-order, and social contract. For example in the obedience / punishment subset there are words like "punished", "awarded", "punishment", "award" and authority / social order contains law-related words like "sentenced", "legal" or "arrested". To examine how emotional and social consequences work together, we created another lexicon, a combination of Kohlberg's theory-based set with the Nakamura's literature-based set. We named the former "Ethical Lexicon", and the latter "Combined Lexicon".

3 Experiments and Results

In this section we introduce experiments we conducted to investigate the effectiveness of our approach in the task of automatic praiseworthy act recognition.

3.1 Input Acts

Web resources used in the study give an opportunity to process any kind of act but this freedom causes difficulties with choosing a fair and balanced input. To deal with this problem we created two sets, one generated automatically and evaluated by subjects, and second one created by the same subjects specifically instructed to give examples of praiseworthy and not praiseworthy acts different from these which they labeled. By introducing these two types we tried to find a balance between "any input" (because the algorithm should recognize neutral acts) and more specific, manually crafted set of correct data.

Automatically Generated Set

For creating the first set we utilized 200 verbs from the Statistical Lexicon with the highest hit number in the blog corpus (100 from positive subset and 100 from negative subset) and paired them with nouns most frequently co-occurring within Japanese Frames dataset automatically generated from the biggest Japanese Web corpus [Kawahara and Kurohashi, 2006]. In order to limit the number of acts and to maintain sufficient coverage (to observe to what extent the automatically polarized words are efficient), we added two conditions. The noun object must be included in the Statistical Lexicon and the generated act must appear at least ten times in the blog corpus. Hence, if e.g. verb "keep" from the lexicon was co-occurring frequently with object noun "promise" and the phrase "to keep a promise" was found more than 10 times in the blog corpus, the phrase was treated as a common human act and became an input. With this method we generated 119 acts which were then evaluated by three judges (one female in her fifties, one male university student and one female secondary school pupil) by labeling the set as *praiseworthy*, *not praiseworthy* or *hard to tell*. The majority vote (three judges agreed or two agreed and the third answered "hard to tell") resulted in 54 acts – 31 worth praising as *tomodachi-o iwau* ("to congratulate a friend") or *chichi-o shitau* ("to admire one's father") and 23 not worth praising as *tanin-o nikumu* ("to hate somebody") or *itami-o shiiru* ("to impose pain upon someone"). Two examples of acts on which agreement was not reached are *hiza-o kussuru* ("to bend one's knees / to yield to someone") and *yami-o kowagaru* ("to be afraid of darkness"). The labeled data became both the input and first correct data set and we named it "Automatically Generated Set".

Manually Created Set

Because the automatically retrieved input set was biased toward Statistical Lexicon we asked the same group of three people to think of acts worth praising and not worth praising. The created set (from now on called "Manually Created Set") contained 64 acts – 32 of praiseworthy ones as *shiken-ni goukaku suru* ("passing an exam") or *tetsudai-o suru* ("helping someone"), and not worth praising as *yakusoku-o mamoranai* ("not to keep a promise"), *kenka-o suru* ("to quarrel / to have a fight"). Differently from the Automatically Generated Set, although the creators have seen examples of acts in the evaluation process, Manually Created Set was not restricted and in consequence included more diverse forms containing not only negations but also adverbs and passive /

Table 1: Results for Automatically Generated Set of input acts.

	Matched / All	Correct
Statistical Lexicon	54 / 54	83.3%
Literature Lexicon	42 / 54	66.7%
Ethical Lexicon	17 / 54	58.8%
Combined Lexicon	45 / 54	68.9%

Table 2: Results for the Manually Created Set of input acts.

	Matched / All	Correct
Statistical Lexicon	52 / 64	63.5%
Literature Lexicon	45 / 64	84.4%
Ethical Lexicon	39 / 64	84.6%
Combined Lexicon	44 / 64	90.9%

double verbs as in *jiko-chuushin-teki ni koudou-o suru* ("to act selfishly") and *iwareta koto-o yaranai* ("not to do what one was told").

3.2 Effectiveness Comparison between Lexicons

Having two sets of acts with their human evaluation prepared, we have performed a series of experiments to examine our system's accuracy when using above described lexicons in the task of recognizing praiseworthy acts.

Statistical Lexicon

Tested with acts from the Automatically Generated Set, the Statistical Lexicon achieved 83.3% of correct recognitions. To confirm our assumption that matching should be performed only on the right side of an act phrase because it is where consequences of the act are usually written (see Figure 2), we have also run additional tests and confirmed that analyzing left sides achieves significantly lower accuracy (66.7%). Matching within the whole sentence did not bring any improvement in results, besides it doubled searching time. Examples of correctly recognized acts are *shouri-o iwau* ("to celebrate victory") and *kenkou-o mamoru* ("to care about one's health"). On the other hand, *tsumi-o kuiru* ("to regret one's sins") or *shi-o kanashimu* ("to grieve one's death") were recognized incorrectly due to noisy polarity in the Statistical Lexicon.

When tested with Manually Created Set, the results of Statistical Lexicon dropped as expected. Left side matching brought only 53.7% correct recognitions while again the right side matching surpassed the left side achieving 63.5% and the whole sentences scored significantly lower (58.2%). All other comparison of results between left side, right side and whole sentences confirmed this trend, therefore, in order to avoid confusion, all remaining results we introduce, are from the matches performed on the right sides following input act phrases.

Literature Lexicon

The Literature Lexicon surpassed much larger Statistical Lexicon when Manually Created Set acts were input but was significantly less accurate with acts from Statistical Lexicon (see Table 1 and Table 2). The perfect recognition rate (54/54 matched) may suggests that if a new, less noisy method for the automatic estimation of word polarity is proposed and it covers all words in every possible input, the Statistical Lexicon would outperform the Literature Lexicon also when fed with acts from Manually Created Set. Nevertheless, it would be very costly and avoiding polarizing neutral words seems to be difficult, hence we believe that using manually crafted, small lexicons is currently more realistic approach for the automatic recognition (and annotation) of praiseworthy acts.

Ethical Lexicon

The smallest of all used lexicons, based on Kohlberg's theory and utilized in automatic ethical recognition task performed worst when the Automatically Generated Set of acts was input but outperformed both Statistical and Literature Lexicons when the Manually Created Set of acts was used.

Combined Lexicon

We managed to confirm that the combination of Ethical and Literature Lexicons performed better than separated ones when the Manually Created Set of acts was used. However, its accuracy was still lower than Statistical Lexicon matching sentences retrieved with the Automatically Created Set of acts.

3.3 Additional Experiments

As we aim at recognizing praiseworthy acts in everyday conversation, the correct recognition of more natural input acts is more important than the correct recognition of less natural input acts. To be sure if Statistical Lexicon could perform better with Manually Created Set we conducted a series of additional tests increasing the range of positive and negative words to see if heuristically chosen size of 3,000 was correct. We examined 10 sizes starting from 500 words size increasing it by 500 each time up to 5,000 and also tested the whole unbalanced list from -1 to 1. It appeared (See Figure 3) that accuracy grows till 1,500 words (increase from 72.9% to 80.8%) but when a larger sets are used, the results start to decrease and never exceed these of the Literature Lexicon (84.4%).

4 Adding Semantic Categories

After analyzing sentences which include praiseworthy act but were not counted due to insufficient number of words in lexicons we decided to examine if we could automatically add some valuable information to other words and see if the information influences the act of praising. We chose semantic categorization and used "Bunrui-Goi-hyo" (Word List by Semantic Principles) [NLRI, 1964] containing 32,600 semantically categorized words collected from 90 contemporary Japanese newspapers. For example the list groups words under categories as "Thoughts / Opinions / Doubts", "Helping / Rescuing" or "Profit / Loss". Our idea was to add simple weighs (count +1) to words that belong to categories which tend to be praiseworthy. In order to examine which categories reveal such tendencies we retrieved from the corpus all sentences containing acts labeled by human subjects

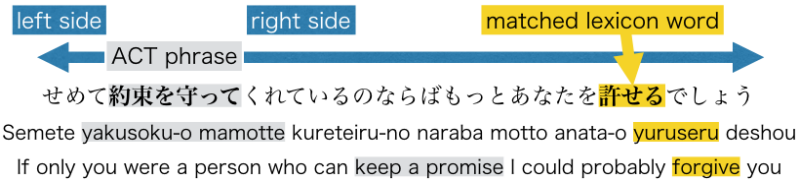


Figure 2: Example sentence from the corpus with input act and a matched Ethical Lexicon word on the right side.

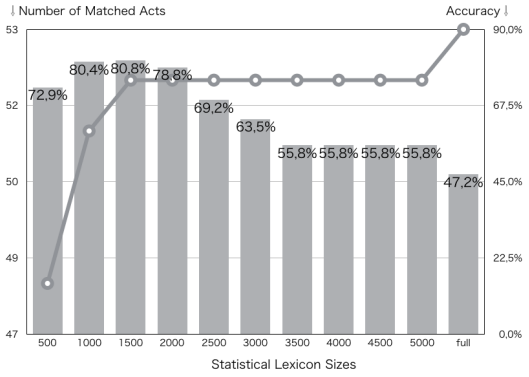


Figure 3: Results of additional experiments for investigating accuracy changes when using different sizes of the Statistical Lexicon.

as praiseworthy and not praiseworthy. Then a simple script counted how many other words in both datasets belong to which semantic category. For example if a blog sentence was "I lost the confidence in myself after he spoke ill about me", the script was adding negative points to categories as "Profit/Loss" (*lost*) or "Thoughts / Opinions / Doubts" (*confidence*). Because some categories contained thousands of words and other only a few, we decided to assign weights according to differences between frequencies. Examples of categories with distinctly different frequencies are shown in Table 3. Then, in order to ease unbalance between sizes of both categories, we experimented with combinations of weight sets and discovered that accuracy is highest for both praiseworthy and not praiseworthy acts when the former uses weights created from group b) and the latter uses c) (refer to Table 3).

4.1 Result Comparison

To see if semantic categorization is effective, we repeated all experiments scoring not only matched lexicon words but also other words that belong to specific categories (those with tendencies to be praiseworthy or not praiseworthy). Because among semantic categories supposedly specific to praiseworthy acts there were ones like Losing and Disappointment, we expected rather low accuracy, but quite surprisingly semantic weighting helped improving all previous results (see Table 4 and Table 5). Even when we excluded lexicon words count entirely, the semantic categories alone achieved slightly better precision than Ethical Lexicon when the Automatically Generated Set of acts was input. The highest precision when Manually Created Set was used increased the precision of Literature and Ethical Lexicons achieving 94%.

5 Conclusion, Future Work and Discussion

In this paper we introduced a simple matching algorithm allowing an agent to recognize human acts worth praising with maximal 94% agreement with human subjects by using lexicons (words sets) and Web resources (a blog corpus). The best results were achieved by Literature and Combined Lexicons with Semantic Categories support when manually created example acts were input. There is still plenty of room for improvement and we plan to increase the coverage of lexicons by matching synonyms, too. We also are experimenting with changing counting method according to adverbs preceding matching phrases ("a little bit sad" could be scored lower than e.g. "so freaking sad"). As the act of praising is very subjective and depends on many factors, we are planning to perform wide, possibly intercultural, surveys. We would like to conclude with underlining a wider importance of the ability to automatically recognize praiseworthy acts by a machine. Recent worries about Artificial Intelligence taking control over their users could be, at least in our opinion, eased by positive examples. Companion robots, while helping at home and e.g. running memory-quizes for users with Alzheimer disease, need to be trusted and gaining the trust will be difficult without sharing similar values. Our common recognition and evaluation of a fellow human's behavior can be measured with shallow sentiment analysis techniques on vast textual data which express our experiences and feelings. The proposed method demonstrates that the noisy Web resources like blogs, when processed carefully, can become one way to equip artificial agents with a human-like capacity of telling right from wrong without leaning to any specific philosophy or religion. We believe that a trustworthy machine should rather operate on estimating overall positive and negative consequences than on methods based on explicit rules decided by one or only few programmers. The proposed system can easily "explain" its decisions by giving examples of retrieved experiences or by presenting a voting ratio, while most of machine learning based methods are "black boxes" and may lead to trust issues. Having said so, we believe that our method could help to automatically annotate data, which is crucial for machine learning.

Table 3: Examples of frequency differences of semantic categories specific to praiseworthy and not praiseworthy acts

Difference	**Praiseworthy acts**
a) More than 4 times:	Helping / Rescuing, Giving / Receiving, Profit / Loss, Winning / Losing, School / Military, Lending / Borrowing, Physiology, Marking / Signing, etc.
b) More than 3 times:	Talents, Planning, Specialist jobs, Associations / Groups, Events / Ceremonies, etc.
c) More than 2 times:	Economy / Income / Expenditure, Formation, Meaning / Problem / Purpose, Desire / Expectance / Disappointment, etc.

Difference	**Not praiseworthy acts**
a) More than 4 times:	Respecting / Thanking / Trusting, Creating / Writing, Old / New / Slow / Fast, Treatment, Graphs / Tables / Formulas, etc.
b) More than 3 times:	Acquisition, Eye / Mouth / Nose functions, Roads / Bridges, Land vehicles, Fear / Anger, etc.
c) More than 2 times:	Linguistic activities, Birds, Associations, Distress / Sorrow, Partners / Colleagues, etc.

Table 4: Effectiveness comparison of implementing semantic categories (Automatically Generated Set).

	Matched / All	Correct
Semantic Category (SC)	52 / 54	78.8%
Statistical Lexicon + SC	54 / 54	85.2%
Literature Lexicon + SC	54 / 54	81.5%
Ethical Lexicon + SC	52 / 54	76.9%
Combined Lexicon + SC	54 / 54	85.2%

Table 5: Effectiveness comparison of implementing semantic categories (Manually Created Set).

	Matched / All	Correct
Semantic Category (SC)	50 / 64	92.0%
Statistical Lexicon + SC	52 / 64	88.5%
Literature Lexicon + SC	50 / 64	94.0%
Ethical Lexicon + SC	50 / 64	90.0%
Combined Lexicon + SC	50 / 64	94.0%

References

[Bandura, 1977] Albert Bandura. Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, 84(2):191, 1977.

[Burnard, 2003] Philip Burnard. Ordinary chat and therapeutic conversation: phatic communication and mental health nursing. *Journal of Psychiatric and Mental Health Nursing*, 10(6):678–682, 2003.

[Cambria *et al.*, 2013] E. Cambria, B. Schuller, Yunqing Xia, and C. Havasi. New avenues in opinion mining and sentiment analysis. *Intelligent Systems, IEEE*, 28(2):15–21, March 2013.

[Carton, 1996] John S Carton. The differential effects of tangible rewards and praise on intrinsic motivation: A comparison of cognitive evaluation theory and operant theory. *The Behavior Analyst*, 19(2):237, 1996.

[Crowell *et al.*, 1988] Charles R Crowell, D Chris Anderson, Dawn M Abel, and Joseph P Sergio. Task clarification, performance feedback, and social praise: Procedures for improving the customer service of bank tellers. *Journal of Applied Behavior Analysis*, 21(1):65–71, 1988.

[Deng and Wiebe, 2014] Lingjia Deng and Janyce Wiebe. Sentiment propagation via implicature constraints. In *EACL*, pages 377–385, 2014.

[Ding and Riloff, 2016] Haibo Ding and Ellen Riloff. Acquiring knowledge of affective events from blogs using label propagation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, 2016.

[Fogg and Nass, 1997] B.J. Fogg and C. Nass. Silicon sycophants: the effects of computers that flatter. *International Journal of Human-Computer Studies*, 46(5):551 – 561, 1997.

[Garland *et al.*, 2008] Ann F Garland, Kristin M Hawley, Lauren Brookman-Frazee, and Michael S Hurlburt. Identifying common elements of evidence-based psychosocial treatments for children's disruptive behavior problems. *Journal of the American Academy of Child & Adolescent Psychiatry*, 47(5):505–514, 2008.

[Henderlong and Lepper, 2002] Jennifer Henderlong and Mark R Lepper. The effects of praise on children's intrinsic motivation: a review and synthesis. *Psychological bulletin*, 128(5):774, 2002.

[Hofmann *et al.*, 2012] Stefan G Hofmann, Anu Asnaani, Imke JJ Vonk, Alice T Sawyer, and Angela Fang. The efficacy of cognitive behavioral therapy: a review of meta-analyses. *Cognitive therapy and research*, 36(5):427–440, 2012.

[Kanouse *et al.*, 1981] David E Kanouse, Peter Gumpert, and Donnah Canavan-Gumpert. The semantics of praise. *New directions in attribution research*, 3:97–115, 1981.

[Kaptein *et al.*, 2010] Maurits Kaptein, Panos Markopoulos, Boris Ruyter, and Emile Aarts. Two acts of social intelligence: the effects of mimicry and social praise on the eval-

uation of an artificial agent. *AI & SOCIETY*, 26(3):261–273, 2010.

[Kawahara and Kurohashi, 2006] Daisuke Kawahara and Sadao Kurohashi. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 176–183, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[Koestner *et al.*, 1987] Richard Koestner, Miron Zuckerman, and Julia Koestner. Praise, involvement, and intrinsic motivation. *Journal of personality and social psychology*, 53(2):383, 1987.

[Kohlberg, 1981] Lawrence Kohlberg. *The Philosophy of Moral Development*. Harper and Row, 1th edition, 1981.

[Lipnevich and Smith, 2008] Anastasiya A Lipnevich and Jeffrey K Smith. Response to assessment feedback: The effects of grades, praise, and source of information. *ETS Research Report Series*, 2008(1):i–57, 2008.

[Nakamura, 1993] Akira Nakamura. *Kanjo hyogen jiten [Dictionary of Emotive Expressions]*. Tokyodo Publishing, 1993.

[NLRI, 1964] National Language Research Institute NLRI. *Bunrui Goi Hyo (Word List by Semantic Principles, in Japanese)*. Shuei Shuppan, 1964.

[Ptaszynski *et al.*, 2012] Michal Ptaszynski, Pawel Dybala, Rafal Rzepka, Kenji Araki, and Yoshio Momouchi. Yacis: A five-billion-word corpus of japanese blogs fully annotated with syntactic and affective information. In *Proceedings of The AISB/IACAP World Congress*, pages 40–49, 2012.

[Rzepka and Araki, 2015] Rafal Rzepka and Kenji Araki. *Rethinking Machine Ethics in the Age of Ubiquitous Technology*, chapter Semantic Analysis of Bloggers Experiences as a Knowledge Source of Average Human Morality, pages 73–95. Hershey: IGI Global, 2015.

[Sarma *et al.*, 2014] Bandita Sarma, Amitava Das, and Rodney D Nielsen. A framework for health behavior change using companionable robots. *INLG 2014*, page 103, 2014.

[Socher *et al.*, 2013] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

[Strain *et al.*, 1983] Phillip S Strain, Deborah L Lambert, Mary Margaret Kerr, Vaughan Stagg, and Donna A Lenkner. Naturalistic assessment of children's compliance to teachers'requests and consequences for compliance. *Journal of Applied Behavior Analysis*, 16(2):243–249, 1983.

[Strapparava and Mihalcea, 2008] Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560. ACM, 2008.

[Swann *et al.*, 1992] William B Swann, Richard M Wenzlaff, and Romin W Tafarodi. Depression and the search for negative evaluations: more evidence of the role of self-verification strivings. *Journal of Abnormal Psychology*, 1992.

[Takamura *et al.*, 2005] Hiroya Takamura, Takashi Inui, and Manabu Okumura. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 133–140. Association for Computational Linguistics, 2005.

[Takeuchi *et al.*, 2010] Koichi Takeuchi, Suguru Tsuchiyama, Masato Moriya, and Yuuki Moriyasu. Construction of argument structure analyzer toward searching same situations and actions. Technical Report 390, IEICE technical report. Natural language understanding and models of communication, jan 2010.

[Tejima and Bunki, 2001] Noriyuki Tejima and Hitomi Bunki. Feasibility of measuring the volition level in elderly patients when using audio encouragement during gait training physical therapy. In *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE*, volume 2, pages 1393–1395. IEEE, 2001.

[Tejima *et al.*, 1998] Noriyuki Tejima, Yoko Takahashi, and Hitomi Bunki. Verbal-encouragement algorithm in gait training for the elderly. In *Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE*, volume 5, pages 2724–2725. IEEE, 1998.

[Weiner *et al.*, 1972] Bernard Weiner, Heinz Heckhausen, and Wulf-Uwe Meyer. Causal ascriptions and achievement behavior: a conceptual analysis of effort and reanalysis of locus of control. *Journal of personality and social psychology*, 21(2):239, 1972.

[Wilson *et al.*, 2005] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

[Wright *et al.*, 2005] Jesse H. Wright, Andrew S. Wright, Anne Marie Albano, Monica R. Basco, L. Jane Goldsmith, Troy Raffield, and Michael W. Otto. Computer-assisted cognitive therapy for depression: Maintaining efficacy while reducing therapist time. *The American Journal of Psychiatry*, 162(6):1158–64, Jun 2005.

[Zimmerman *et al.*, 2009] Frederick J Zimmerman, Jill Gilkerson, Jeffrey A Richards, Dimitri A Christakis, Dongxin Xu, Sharmistha Gray, and Umit Yapanel. Teaching by listening: The importance of adult-child conversations to language development. *Pediatrics*, 124(1):342–349, 2009.

Multimodal Sentiment Analysis of Telugu Songs

Harika Abburi, Eswar Sai Akhil Akkireddy, Suryakanth V Gangashetty, Radhika Mamidi
Langauage Technology Research Center
IIIT Hyderabad India
{harika.abburi, eswarsai.akhil}@research.iiit.ac.in
{svg, radhika.mamidi}@iiit.ac.in

Abstract

In this paper, an approach to detect the sentiment of a song based on its multi-modality natures (text and audio) is presented. The textual lyric features are extracted from the bag of words. By using these features, Doc2Vec will generate a single vector for each song. Support Vector Machine (SVM), Naive Bayes (NB) and a combination of both these classifiers are developed to classify the sentiment using the textual lyric features. Audio features are used as an add-on to the lyrical ones which include prosody features, temporal features, spectral features, tempo and chroma features. Gaussian Mixture Models (GMM), SVM and a combination of both these classifiers are developed to classify the sentiment using audio features. GMM are known for capturing the distribution in the features and SVM are known for discriminating the features. Hence these models are combined to improve the performance of sentiment analysis. Performance is further improved by combining the text and audio feature domains. These text and audio features are extracted at the beginning, ending and for the whole song. From our experimental results, it is observed that the first 30 seconds(s) of a song gives better performance for detecting the sentiment of the song rather than the last 30s or from the whole song.

1 Introduction

Sentiment analysis is defined as a task of finding the opinion about specific entities. In our case it is a task of finding the sentiment of a song. With the growing amount of music and the demand of human to access the music information retrieval, music sentiment analysis is emerging as an important and essential task for various system and applications. To extract the sentiment, thousands of text, audio and video documents will process in few seconds . Sentiment analysis mainly focuses on two approaches, text based and audio based [Tyagi and Chandra, 2015]. For any approach sentiment can be extracted using sentiment classification techniques like machine learning approach, lexicon based approach and hybrid approach [Medhat *et al.*, 2014].

In lyric-based song sentiment classification, sentiment-vector space model is used for song sentiment classification [Xia *et al.*, 2008]. Experiments are done on two approaches: knowledge-based and machine learning. In knowledge-based, HowNet [Dong *et al.*, 2010] is used to detect the sentiment words and to locate the sentiment units with in the song lyric. In machine learning, the SVM algorithm is implemented based on Vector Space Model (VSM) and sentiment-Vector Space Model (s-VSM), respectively. Experiments show that s-VSM gives better results compared to VSM and knowledge-based. A previous work includes sentiment analysis for mining the topics from songs based on their moods [Shanmugapriya and Dr.B.Srinivasan, 2015]. The input lyrics files are measured based on the wordnet graph representation and the sentiments of each song are mined using Hidden Markov Model (HMM). Based on single adjective words available from the audio dataset USPOP, a new dataset is derived from the last.fm tags [Hu *et al.*, 2007]. Using this dataset, K-means clustering method is applied to create a meaningful cluster-based set of high-level mood categories for music mood classification. This set was not adopted by others because mood categories developed by them were seen as a domain oversimplification. The authors in [Hu *et al.*, 2009] presented the usefulness of text features in music mood classification on 18 mood categories derived from user tags and they show that these text features outperform audio features in categories where samples are more sparse. An unsupervised method to classify music by mood is proposed in [Patra *et al.*, 2013]. Fuzzy c-means classifier is used to do the automatic mood classification.

In audio-based song sentiment classification: A method is presented for audio sentiment detection based on KeyWord Spotting (KWS) rather than using Automatic Speech Recognition (ASR) [Kaushik *et al.*, 2015]. Experiments show that the presented method outperform the traditional ASR approach by 12 percent increase in classification accuracy. Another method for detecting the sentiment from natural audio streams is presented [Kaushik *et al.*, 2013]. To obtain the transcripts from the video, ASR is used. Then a sentiment detection system based on Maximum Entropy modeling and Part of Speech tagging is used to measure the sentiment of the transcript. The approach shows that it is possible to automatically detect sentiment in natural spontaneous audio with good accuracy. Instead of using KWS and ASR we can di-

Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016, pages 48-52, New York City, USA, July 10, 2016.

rectly extract the features like prosody, spectral etc to detect the sentiment of a song from audio. For music audio classification, instead of using Mel Frequency Cepstral Coefficients (MFCC) and chroma features separately combination of both gives better performance. Because chroma features are less informative for classes such as artist, but contain information which is independent of the spectral features [Ellis, 2007]. Due to this reason in our work, experiments are done by combining both features along with some other features.

Instead of using only lyrics or only audio, research is also done on combinations of both the domains. In [Hu and Downie, 2010] work is done on the mood classification in music digital libraries by combining lyrics and audio features and discovered that complementing audio with lyrics could reduce the number of training samples required to achieve the same or better performance than single source-based systems. Music sentiment classification using both lyrics and audio is presented [Zhong *et al.*, 2012]. For lyric sentiment classification task, CHI approach and an improved difference-based CHI approach were developed to extract discriminative affective words from lyrics text. Difference-based CHI approach gives good results compare to CHI approach. For audio sentiment classification task, features like chroma, spectral etc. are used to build SVM classifier. Experiments show that the fusion approach using data sources help to improve music sentiment classification. In [Jamdar *et al.*, 2015], [Wang *et al.*, 2009] music is retrieved based on both lyrics and melody information. For lyrics, keyword spotting is used and for melody MFCC and Pitch features are extracted. Experiments show that by combining both modalities the performance is increased.

In this work, a method to combine both lyrics and audio features is explored for sentiment analysis of songs. As of now, less research is done on multimodal classification of songs in Indian languages. Our proposed system is implemented on Telugu database. For lyrics, Doc2Vec is used to extract the fixed dimension feature vectors of each song. SVM and Naive Bayes classifiers are built to detect the sentiment of a song due to their excellence in text classification task. For audio, several features are extracted like prosody, temporal, spectral, chroma, harmonics and tempo. Classifiers that are built to detect the sentiment of a song are SVM, GMM and combination of both. It is observed that in the literature a lot of work is done on whole song to know the sentiment, but the whole song will not give good accuracy because the whole song may or may not carry the same attribute like happy (positive) and sad (negative). The beginning and the ending parts of the song includes the main attribute of that song. Hence, experiments are done on different parts of the song to extract the sentiment.

The rest of the paper is organized as follows:Database and classifiers used in this work is discussed in section 2 and sentiment analysis using lyric features is discussed in section 3. Sentiment analysis using audio features is discussed in section 4. Multimodal sentiment analysis and experimental results in proposed method for detecting the sentiment of a song is discussed in section 5. Finally, section 6 concludes the paper with a mention on the future scope of the present work.

2 Database and Classifiers used in this study

The database used in this paper is collected from the Youtube which is a publicly available source. A total of 300 Telugu movie songs and lyrics corresponding to each song are taken. The two basic sentiments presented in the database are: Happy and Sad. Joyful, thrilled, powerful, etc are taken as happy sentiment and ignored, depressed, worry, etc are taken as sad sentiment. As our native language is Telugu, work is implemented on Telugu songs which don't have any special features compared to other language songs. Telugu songs are one of the popular categories of Indian songs and are present in Tollywood movies. Most of the people belonging to the south part of India will listen to these songs.The songs include variety of instruments along with the vocals. Here the main challenging issue is the diversity of instruments and vocals. The average length of each song is three minutes thirty seconds and average number of words in lyrics for each song is around 300. The database is annotated for the sentiment happy and sad by three people. Annotators are provided with the two modalities such as text and audio to correctly figure out the sentiment of a song. Then based on inter-annotator agreement, 50 happy songs and 50 sad songs are selected because some songs seems to be happy or sad for one annotator and neutral to another annotator. So, only 100 songs are selected out of 300. Inter-annotator agreement is a measure of how well two or more annotators can make the same annotation decision for a certain category. Among them 40% of songs are used for training and 60% of songs are used for testing.

2.1 Naive Bayes

Naive Bayes classifier is a probabilistic classifier of words based on the Bayes theorem with an independence assumption that words are conditionally independent of each other. This assumption does not affect the accuracy in text classification but makes really fast classification algorithm. Despite the assumptions that this technique uses, Naive Bayes performs well in many complex real-world problems. Multinomial Naive Bayes is used in our system where the multiple occurrences of the words matter a lot in the classification problem.

The main theoretical drawback of Naive Bayes method is that it assumes conditional independence among the linguistic features. If the main features are the tokens extracted from texts, it is evident that they cannot be considered as independent, since words co-occurring in a text are somehow linked by different types of syntactic and semantic dependencies. Despite its simplicity and conditional independence assumption, Naive Bayes still tends to perform surprisingly well [Rish, 2001]. On the other hand, more sophisticated algorithms might yield better results; such as SVM.

2.2 Support Vector Machines

Support vector machine classifier is intended to solve two class classification problems. The basic principle implemented in a support vector machine is that the input vectors which are not linearly separable are transformed to a higher dimensional space and an optimum liner hyperplane is designed to classify both the classes. An SVM [Campbel *et al.*,

2006] is a two-class classifier constructed from sums of a kernel functions.

2.3 Gaussian Mixture Models

GMMs are well known to capture the distribution of data in the feature space. A Gaussian mixture density is a sum of M weighted component densities [Reynolds and Rose, 1995] given by the equation:

$$p(x_k|\lambda) \quad \sum_{r=1}^{M} w_r K_r(x_k) \tag{1}$$

where x_k is an N dimensional input vector,
$K_r(x_k), r \quad 1...M$ are the component densities and
$w_r, r \quad 1...M$ are the weights of the mixtures.

The product of the component Gaussian with its mixture weight i.e., $K_p(x_k)w_r$ is termed as component density. Sum of the component densities is given by Gaussian mixture density. The accuracy in capturing the true distribution of data depends on various parameters such as dimension of feature vectors, number of feature vectors and number of mixture components. In this work expectation maximization (EM) algorithm is used to train the GMM models using audio features.

3 Sentiment Analysis using Lyric Features

This section describes the process of extracting the textual lyrics of a song. These features are then used to build a classifier of positive or negative sentiment of a song. In Preprocessing step, lyrics which contain stanza names like "pallavi" and "charanam" were removed because, as the lyrics are collected from the Internet the headings ("pallavi" and "charanam") are common for each song which does not act like a feature to detect the sentiment of the song. If the same line has to repeated, it is represented as "x2" in the original lyrics, so "x2" is removed and the line opposite to that is considered as twice. For each song in a database one feature vector with 300 dimension is generated for better results. As we have 100 files, 100 feature vectors are generated one for each song. For checking the accuracy, each song is manually annotated and is given a tag like happy or sad.

Here Doc2Vec model is used for associating random documents with labels. Doc2vec modifies word2vec algorithm to a unsupervised learning of continuous representations for larger blocks of text such as sentences, paragraphs or whole documents means Doc2vec learns to correlate labels and words rather than words with other words. In the word2vec architecture, the two algorithms used are continuous bag of words and skip-gram and for the doc2vec architecture, the corresponding algorithms are distributed memory and distributed bag of words. All songs are given as input to the doc2vec. This generates a single vector that represents the meaning of a document, which can then be used as input to a supervised machine learning algorithm to associate documents with labels. Song sentiment analysis based on lyrics can be viewed as a text classification task which can be handle by SVM and NaiveBayes (NB) algorithms due to their better classification performance. Both SVM and NaiveBayes classifiers are trained with vectors generated from the doc2vec. After calculating the probabilities from both the classifiers, average probabilities of them is computed. Which ever class gives highest average probability that test case is hypothesized from that class. Like this these two classifiers are compared. By combining both the classifiers, rate of detecting the sentiment of a song is improved. Given a test data song, the trained models classifies it as either happy or sad. Three experiments are done on each song:beginning 30 seconds, last 30 seconds and for the whole song.

Table 1: Sentiment Classification with Lyric Features

	SVM	NB	SVM+NB
Whole song	60.6	52.3	70.2
Beginning of a song	67.5	57.3	**75.7**
Ending of a song	64.4	55.8	72.4

From Table 1 it is observed that a combination of both the classifiers gives high percentage for beginning of the song compared to the ending and whole song. Whole song gives less accuracy in detecting the sentiment of a song. By keeping the training data set constant several experiments are done on the test data. The average performance of sentiment analysis for beginning, ending and for whole song is 75.7, 72.4 and 70.2 respectively.

4 Sentiment Analysis using Audio Features

This section describes the process of extracting the audio features of a song. These features are then used to build a classifier of positive or negative sentiment of a song. Each song underwent the preprocessing step of converting mp3 files into wave file (.wav format), into 16 bit, 16000 Hz sampling frequency and to a mono channel. To extract a set of audio features like mfcc, chroma, prosody, temporal, spectrum, harmonics and tempo from a wave file openEAR/openSMILE toolkit [Eyben *et al.*, 2010] is used. Brief details about audio features are mentioned below:

- Prosody features include intensity, loudness and pitch that describe the speech signal.
- Temporal features also called as time domain features which are simple to extract like the energy of signal, zero crossing rate.
- Spectral features also called as frequency domain features which are extracted by converting the time domain into frequency domain using the Fourier Transform. It include features like fundamental frequency, spectral centroid, spectral flux, spectral roll-off, spectral kurtosis, spectral skewness. These features can be used to identify the notes, pitch, rhythm, and melody.
- In Mel-frequency Cepstral Coefficients (MFCC) (13 dimension feature vector) the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely.

- Chroma features (12 dimension feature vector) are most popular feature in music and is extensively used for chord, key recognition and segmentation.
- Harmonic tempo is the rate at which the chords change in the musical composition in relation to the rate of notes.

Although this toolkit is designed for the emotion recognition, the research has been done on sentimental analysis by using the same toolkit which is succeeded [Mairesse *et al.*, 2012]. As prosody have been used before for the task of emotion recognition in speech, it has also been experimented for the task of sentiment analysis by the authors [Mairesse *et al.*, 2012]. Three experiments are performed here:beginning 30 seconds, last 30 seconds and for the whole song. Features that are extracted are trained on the classifiers such as SVM, GMM and combination of both. GMM are known for capturing the distribution in the features and SVM are known for discriminating the features. Hence these models are combined improve the performance of detecting the sentiment of a song using the audio features. GMM need more features for training compared to Naive Bayes and SVM, but in textual part we have less features (only one feature vector for one song using doc2vec). Where as for audio, several features are their because for each song features are extracted at frame level with a frame size of 20 ms. So for acoustic models GMM and SVM are used where as for linguistic features Naive Bayes and SVM are used. A total of 40 dimension feature vectors are extracted, each of them obtained at frame level. During the feature extraction, frame size of 25ms and frame shift of 10ms are used. In this work, number of mixtures for GMM models (64) and Gaussian kernel parameters for SVM models are determined empirically.

Table 2: Sentiment Classification with Audio Features

	SVM	GMM	SVM+GMM
Whole song	52.8	54.9	69.7
Beginning of the song	55.8	73.5	**88.3**
Ending of the song	64.7	61.7	82.35

From Table 2 it is observed that the whole song gives less performance in detecting the sentiment of a song because the whole song will carries different attributes (happy and sad) which is not clear. So by using part of song, the performance is increased. Hence experiments are done even on beginning and ending of the song. Combination of both classifiers gives a high percentage for beginning of the song compared to the ending of the song. SVM is best performed at the ending of the song, GMM is best performed at the beginning of the song. By keeping training data set constant several experiments are done on the test data. The average performance of sentiment analysis for beginning, ending and for whole song is 88.3, 82.3 and 69.7 respectively.

5 Multimodal Sentiment Analysis

The main advantage that comes with the analysis of audio as compared to their textual data is it will have voice modularity. In textual data, the only source that we have is information regarding the words and their dependencies, which may sometime be insufficient to convey the exact sentiment of the song. Instead, audio data contain multiple modalities like acoustic, and linguistic streams. From our experiments it is observed that textual data gives less percentage than the audio, so the simultaneous use of these two modalities will help to create a better sentiment analysis model to detect whether the song is happy or sad.

Sequence of steps in proposed approach is presented in the Figure 1. Table 3 presents the accuracy of sentiment by combining lyrics and audio features. The whole song may not convey sentiment, so there will be lot of similarity between sad and happy features. Hence features extracted from different parts of a song are used to identify the sentiment of the song. To handle the similarity of sentiment classes, decision from different classification models trained using different modalities are combined. By combining both the modalities performance is improved by 3 to 5%.

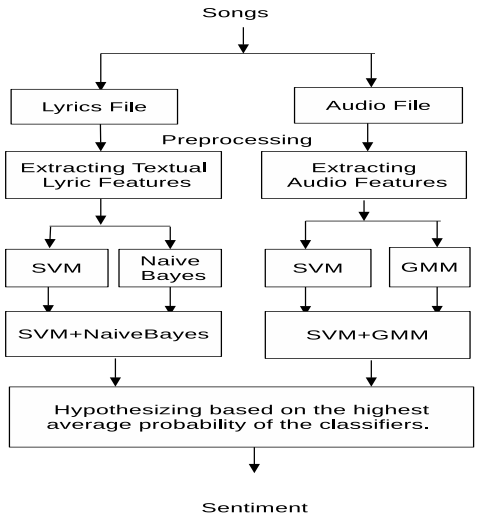


Figure 1: Block diagram of multimodal sentiment analysis of songs

Table 3: Sentiment Classification with Lyric and Audio Features

	Lyric	Audio	Lyric+Audio
Whole song	70.2	69.7	75.8
Beginning of a song	75.7	88.3	**91.2**
Ending of a song	72.4	82.3	85.6

6 Summary and Conclusions

In this paper, an approach to extract the sentiment of a song using both lyrics and audio information is demonstrated.

Lyric features which are generated using Doc2Vec and most efficient audio features like spectral, chroma, etc are used to built the classifiers. Sentiment analysis systems are built using the whole song, beginning of the song and ending of the song. By taking the whole song the performance is very less because the full song will contain more information (features) which is confusing. Hence experiments are done on the beginning and the ending of the songs which are giving better results. Features are extracted from beginning of the song are observed to be giving better performance compared to the whole song and the ending of the song. Because the instruments and vocals which convey the sentiment for beginning part of the song may or may not sustain throughout the song. Several experiments are done by keeping training data constant. The proposed method is evaluated using 100 songs. From the experimental results, recognition rate is observed to be in between 85% to 91.2%. This work can be extended by including more attributes like angry, fear and by extracting more features like rhythm and tonality. The percentage of lyric sentiment analysis can be improved by using rule based and linguistic approach.

References

[Campbel *et al.*, 2006] M William Campbel, P Joseph Cambell, A Douglas Reynolds, Elliot Singer, and A Pedro Torres-Carrasquillo. Support vector machines for speaker and language recognition. *Computer Speech & Language*, 20(2):210–229, 2006.

[Dong *et al.*, 2010] Zhendong Dong, Qiang Dong, and Changling Hao. Hownet and its computation of meaning. In *Proc. 23rd international conference on computational linguistics: demonstrations, association for computational linguistic*, pages 53–56, 2010.

[Ellis, 2007] D. P. W. Ellis. Clasifying music audio with timbral and chroma features. In *Proc. 8th Int. Conf. Music Inf. Retrieval (ISMIR)*, pages 339–340, 2007.

[Eyben *et al.*, 2010] F. Eyben, M. Wollmer, and B. Schulle. opensmile the munich versatile and fast open-source audio feature extractor. In *Proc. ACM Multimedia (MM)*, pages 1459–1462, 2010.

[Hu and Downie, 2010] X. Hu and J. S. Downie. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proc. Joint Conference on Digital Libraries, (JCDL)*, pages 159–168, 2010.

[Hu *et al.*, 2007] X. Hu, M. Bay, and J. S. Downie. Creating a simplified music mood classification ground-truth set. In *Proc. 8th International Conference on Music Information Retrieval*, 2007.

[Hu *et al.*, 2009] Xiao Hu, J. Stephen Downie, and Andreas F. Ehmann. Lyric text mining in music mood classification. In *Proc. 10th International Conference on Music Information Retrieval (ISMIR)*, pages 411–416, 2009.

[Jamdar *et al.*, 2015] Adit Jamdar, Jessica Abraham, Karishma Khanna, and Rahul Dubey. Emotion analysis of songs based on lyrical and audio features. *International Journal of Artificial Intelligence and Applications(IJAIA)*, 6(3):35–50, 2015.

[Kaushik *et al.*, 2013] Lakshmish Kaushik, Abhijeet Sangwan, and John H L. Hansen. Sentiment extraction from natural audio streams. In *proc. ICASSP*, pages 8485–8489, 2013.

[Kaushik *et al.*, 2015] Lakshmish Kaushik, Abhijeet Sangwan, and John H.L. Hansen. Automatic audio sentiment extraction using keyword spotting. In *Proc. INTERSPEECH*, pages 2709–2713, September 2015.

[Mairesse *et al.*, 2012] F. Mairesse, J. Polifroni, and G. Di Fabbrizio. Can prosody inform sentiment analysis? experiments on short spoken reviews. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 5093–5096, 2012.

[Medhat *et al.*, 2014] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering journal*, pages 1093–1113, 2014.

[Patra *et al.*, 2013] B. G. Patra, D. Das, and S. Bandyopadhyay. Unsupervised approach to hindi music mood classification. In *Mining Intelligence and Knowledge Exploration (MIKE 2013), R. Prasath and T. Kathirvalavakumar (Eds.):LNAI 8284*, pages 62–69, 2013. Springer International Publishing.

[Reynolds and Rose, 1995] A Douglas Reynolds and C Richard Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, 1995.

[Rish, 2001] Irina Rish. An empirical study of the naive bayes classifier. In *Proc. IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence*, 2001.

[Shanmugapriya and Dr.B.Srinivasan, 2015] K.P Shanmugapriya and Dr.B.Srinivasan. An efficient method for determining sentiment from song lyrics based on wordnet representation using hmm. *International Journal of Innovative Research in Computer and Communication Engineering*, 3(2):1139–1145, February 2015.

[Tyagi and Chandra, 2015] Atul Tyagi and Nidhi Chandra. An introduction to the world of sentimnt analysis. In *Proc. 28th IRF International Conference*, June 2015.

[Wang *et al.*, 2009] Tao Wang, DongJu Kim, KwangSeok Hong, and JehSeon Youn. Music information retrieval system using lyrics and melody information. In *proc. Asia-Pacific Conference on Information Processing*, pages 601–604, 2009.

[Xia *et al.*, 2008] Yunqing Xia, Linlin Wang, Kam-Fai Wong, and Mingxing Xu. Sentiment vector space model for lyric-based song sentiment classification. In *proc. ACL-08:HLT, Short Papers*, pages 133–136, 2008.

[Zhong *et al.*, 2012] Jiang Zhong, Yifeng Cheng, Siyuan Yang, and Luosheng Wen. Music sentiment classification integrating audio with lyrics. *Information and Computational Science*, 9(1):35–54, 2012.

Author Index