EMTC: Multilabel Corpus in Movie Domain for Emotion Analysis in Conversational Text

Phan Duc-Anh and Yuji Matsumoto
Computational Linguistics Laboratory, Graduate school of Information Science
Nara Institute of Science and Technology, Japan
{phan.duc_anh.oq3, matsu}@is.naist.jp

Abstract

It is proved that in text-based communication such as sms, messengers applications, misinterpretation of partner's emotions are pretty common. In order to tackle this problem, we propose a new multilabel corpus named Emotional Movie Transcript Corpus (EMTC). Unlike most of the existing emotion corpora that are collected from Twitters and use hashtags labels, our corpus includes conversations from movie with more than 2.1 millions utterances which are partly annotated by ourselves and independent annotators. To our intuition, conversations from movies are closer to real-life settings and emotionally richer. We believe that a corpus like EMTC will greatly benefit the development and evaluation of emotion analysis systems and improve their ability to express and interpret emotions in text-based communication.

Keywords: emotion corpus, movie transcript, text-based communication, emotion analysis, multilabel

1. Introduction

In the recent years, we experience rapid development of online communication with the help of mediated devices such as smart phones, tablets, computer. People use talk-over-internet, video conferencing features for both daily and business tasks. Text based methods like emails or text messengers are still very convenient and indispensable to us because of its unique advantages: they do not require intermediate responses and can be used for the sake of record-keeping. However, it is already proved that in any online communication methods, users experience more difficulties in interpreting and conveying emotions than face-to-face communication due to the limitation in communication modality. Furthermore, text-based methods are where difficulties are encountered the most (Kruger et al., 2005; Arimoto and Okanoya, 2016). Therefore, we targeted the effort to build an emotion analysis system focusing on text data. The starting point is to develop an emotional corpus that has conversational texts and is as close to real-life communication as possible.

Existing emotional text corpora are often collected from micro-blog platforms using multiclass scheme - *one emotion per example* (Liew et al., 2016). Most of them are automatically annotated by extracting hashtags rather than by human judgements (Dini and Bittar, 2016; Li et al., 2016). While the text data from micro-blog platforms like Twitters are very convenient and easy to collect, the fact that they are limited in the number of characters (140 for a tweet) differs themselves from daily conversation text and therefore, have limited use in a real-life settings. On the other hand, multiclass scheme has its own limitation. One input is only associated to one emotion. However, it is observed in some research (Liew et al., 2016) that multilabel scheme *with no limitation in the number of emotions per example* is a better and more natural way of annotating emotion labels.

We describe our efforts to construct and annotate partly the Emotional Movie Transcript Corpus (EMTC). Most of the corpus are unsupervised data. We annotated by ourselves 10,000 utterances and use them for training. Finally, the testing data, which include 1000 utterances, are annotated by 5 independent annotators. To our understanding, EMTC is the only emotional corpus that is annotated using multilabel scheme and has conversational text instead of short text like tweets or news headlines (Strapparava and Mihalcea, 2007; Mohammad, 2012b). Moreover, EMTC provide the annotators with movie clips instead of just text to help them give better annotation. Our contributions are summarized as follow:

- We explain the multilabel annotating scheme following Plutchik's theory of emotions (Plutchik, 2001). We then later conclude that our annotating scheme provide much better inter-annotators agreement score than other corpora.
- We present and describe the characteristics of the conversational corpus and the statistics of the annotated data.
- We conduct supervised machine learning experiments to evaluate the emotion classification using our corpus and the word-embedding extracted from it.

2. Related Works

There have been numerous works on building emotion corpora. The first notable work is the ISEAR dataset (Scherer and Wallbott, 1994) which has more than 7,000 responses from participants. The participants are asked to describe the situation where they experience some certain emotions. In our work, we use this dataset to extract collocation features for the manual feature extration step described in the below section. Another corpus is the Semeval-2007 task 14: Affective text (Strapparava and Mihalcea, 2007) which consists of 1,250 news headlines with six Ekman's emotion labels. More recent works are (Mohammad, 2012b; Liew et al., 2016; Dini and Bittar, 2016) where they collect data from micro-blog platforms and automatically annotate

them using hashtags with or without human revision afterwards. The limitation with those corpora is that they only consist of short, independent pieces of text and undoubtedly not close to real-life conversation.

As a matter of fact, modeling emotions in a conversation is indeed a difficult but rewarding task with a wide range of applications. A good system should consider every word in the conversation, the grammatical structure and syntactic variables such as negations, embedded sentences, and type of sentence (question, exclamation, command, or statement), the general context of the conversation, each and every utterances in the conversation - especially when what is said in the previous utterance can have an impact on the emotions of the later one (Collier, 2014). Maybe, because of this complicated nature of the problem, there is a lack of emotional conversation corpus.

Another problem with the existing corpora is the annotating scheme: many works limit the emotion labels to a small number (Mohammad, 2012b; Wang et al., 2015) or only allow annotators to label one emotion per utterance (Yang et al., 2007; Hasegawa et al., 2013). As pointed out in many psychology research (Plutchik, 2001; Russell, 2003), emotions are not mutually exclusive. In fact, in many cases, people may experience a mixture of various emotions at the same time (Choe et al., 2013). Therefore, the corpus for any emotion analysis task should be multilabel. Limiting the number of emotion labels may narrow down the problem but can cause troubles for the annotators to provide correct judgement when the emotions in an example are sophisticated or expressed implicitly.

In our work, we employ Plutchik's theory of emotions and extend the set of labels to a total of 48 labels to provide more freedom to the annotators. The extension and Plutchik's theory will be explained in more details in Section 3., where we present the construction of the corpus. Section 4. investigates the characteristics of our newly built corpus and Section 5. discusses the experiments and evaluation of the corpus. Lastly, Section 6. gives conclusions and future work.

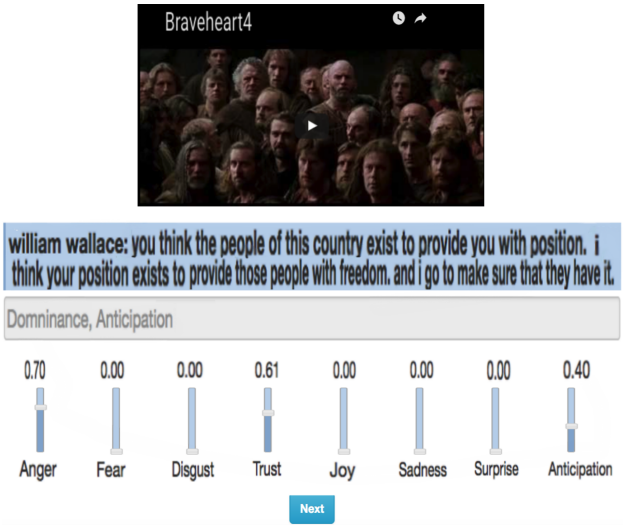
3. Methodology

3.1. Imdb quotes dataset

In order to mimic real-life conversation settings, we rely on the Imdb datasets [1], in particularly, the movie quotes dataset. This dataset includes in total 2,107,863 utterances (turns in conversation) out of 117,425 movies and tv series of all genres such as: thrillers, action, romantic, etc. To our assumption, movies conversation should be close to real-life settings and emotionally rich. We can also easily eliminate the low inter-annotators agreement score problem that is often encountered in other corpus (Strapparava and Mihalcea, 2007; Dini and Bittar, 2016) by providing them the clips from the movies in addition to the transcripts (Figure 1a).

At first, we would also want to measure the judgement of emotion intensity from the annotators, hence the bar measurement. However, the collected numbers are very diverse and unreliable. This is due to disagreement among the annotators and their different interpretation of the emotion intensity during the annotation sessions.

There is also a concern that different culture will give different interpretation of emotion expressions in those provided clips. However, nowadays, as people from different cultures are more exposed to and have more opportunity to watch American movies, they steadily learn how to interpret the emotions from other cultures better (Hareli et al., 2015; Lim, 2016).



(a) UI of the annotating website. Users can choose the appropriate emotions by adjusting the confidence bars or by typing the emotions or dyads into the text box. The dyads are then decomposed automatically into primary emotions and the bars are readjusted

robert the bruce: my hate will die with you.	{"Anger":"0.37","Disgust":"0.40"}
princess isabelle: the king desires peace.	{"Trust":"0.22"}
william wallace: longshanks desires peace?	{"Anger":"0.26","Disgust":"0.34","Surprise":"0.36"...
william wallace: go back to england and tell them ...	{"Anger":"0.52","Disgust":"0.38","Trust":"0.34"}

(b) Examples of annotated transcripts from movie: Brave Heart (1995) -

Figure 1: Annotating scheme of the testing data. Each utterance is annotated with primary emotions

3.2. Plutchik's theory of emotions

The reason for most research to limit the number of emotion categories is to have a better inter-annotators agreement score. The more categories are allowed, the lower the score it becomes. However, by limiting the number of categories, they also limit the freedom of the annotators to give accurate judgements because emotions are sophisticated and the basic emotions can hardly cover all the cases. In our work, we found out a way to avoid this trade-off: Plutchik's theory of emotions.

According to Plutchik, there are eight primary emotions grouped on a positive or negative basis: joy versus sadness; anger versus fear; trust versus disgust; and surprise versus anticipation. Some emotions are similar to the primary ones but different in intensity (Table 1). Some pri-

[1] The datasets are available from http://www.imdb.com/interfaces

mary emotions can be mixed to form more complex emotions 2. Implementing the theory, we allow the annotators to use the full 48 emotion labels from the two tables, the system will automatically decompose the annotated labels into primary emotions later (Figure 1).

Intense emotion	Ecstasy	Admiration	Terror	Amazement
Mild emotion	Serenity	Acceptance	Apprehension	Distraction
Primary emotion	Joy	Trust	Fear	Surprise
Primary opposite	Sadness	Disgust	Anger	Anticipation
Mild opposite	Pensiveness	Boredom	Annoyance	Interest
Intense opposite	Grief	Loathing	Rage	Vigilance

Table 1: Emotions and the opposite of them

Human feelings	**Parent Emotions**	**Opposite feelings**	**Parent Emotions**
Optimism	Anticipation, Joy	Disapproval	Surprise, Sadness
Hope	Anticipation, Trust	Unbelief	Surprise, Disgust
Anxiety	Anticipation, Fear	Outrage	Surprise, Anger
Love	Joy, Trust	Remorse	Sadness, Disgust
Guilt	Joy, Fear	Envy	Sadness, Anger
Delight	Joy, Surprise	Pessimism	Sadness, Anticipation
Submission	Trust, Fear	Contempt	Disgust, Anger
Curiosity	Trust, Surprise	Cynicism	Disgust, Anticipation
Sentimentality	Trust, Sadness	Morbidness	Disgust, Joy
Awe	Fear, Surprise	Aggressiveness	Anger, Anticipation
Despair	Fear, Sadness	Pride	Anger, Joy
Shame	Fear, Disgust	Dominance	Anger, Trust

Table 2: Dyads - Combinations of emotions: two primary emotions can blend together to form another complex one

3.3. Annotation Scheme

To produce labeled data, annotators are asked to watch the corresponding movies with subtitles (Figure 1a) and follow the annotation scheme shown below:

- One utterance may hold zero, one or more emotions at the same time. In case an utterance holds no emotion, it should be annotated with "None." The intensity of emotions is also considered in the labeling phrase (Figure 1).
- The annotators can choose appropriate emotion labels from the list of 48 emotions in tables 1 and 2. The system will decompose the dyads into primary emotions automatically.
- The annotators need to assign the whole utterance which may have two or more sentences with a set of all emotions expressed inside it. There may be cases where conflict emotions according to Plutchik's theory to appear simultaneously in the same utterance as in the last example of the subfigure 1b.

4. Characteristics, the inter-annotators agreement and the word-embedding of the corpus

This corpus includes in total 2,107,863 utterances with 26 millions words, 181,276 of which are unique terms. As mentioned in the above sections, we can only annotate the corpus partly. There are 10,000 utterances that are annotated by the authors ourselves as the training data. The average emotion labels per utterances are 1.68. The testing data are reviewed by 5 independent annotators to form a gold standard data (with majority rule) with 1,000 utterances. The reason for two different datasets was because the annotation sessions are expensive and time consuming. We have to provide the annotators with clips cutting from real movies and match the text from the corpus to the correct scenes in the full movies.

The average labels per utterances are 1.41. We report the inter-annotators agreement score of our testing data in the Table 3 where the performance of each annotator is compared to the gold standard data as ground-truth.

Emotion class	**Accuracy**
Anger	0.72
Fear	0.673
Disgust	0.624
Trust	0.65
Joy	0.606
Sadness	0.584
Surprise	0.575
Anticipation	0.491
Average accuracy (by class)	0.615
Average accuracy (by annotator)	0.43
Average F1 (by annotator)	0.626
Total No. utterances	1,000

Table 3: Inter-annotator Agreement score with gold standard data as ground-truth.

From on the table, it can be concluded that: Our corpus, even when being annotated using multilabel scheme, yields better agreement score to the multiclass - Twitter Emotion Corpus (Mohammad, 2012b) (Average F1-score is 43.7).

We believe that our choice of using a movie corpus and providing movie clips to support the annotation process plays an important factor here.

4.1. Word-embedding of the corpus

Word-embedding is the vector multi-dimensional representations of every words in the corpus. It can be a simple yet effective input features for many machine learning methods. In this research, we follow this approach and create the embedding with 100 dimensions using word2vec (Řehůřek and Sojka, 2010). Table 4 shows the top 5 most similar terms to the primary emotion words and some of the dyads.

Interesting points can be observed from the table: 1) We notice some of the dyads appear in the top similar list of the parent emotions, which - to some extend - validate Plutchik's theory. 2) Most of the top similar terms are quite on point and reasonable. 3) Opposite emotions and dyads sometime appear together, it maybe interesting to investigate the correlation among labels to see if the same phenomenon occurs.

Emotions	Top similar
Anger	Rage, Pain, Hatred, Guilt, Grief
Fear	Hatred, Darkness, Despair, Grief, Desire
Trust	Betray, Confuse, Respect, Underestimate, Threaten
Disgust	Horrified, Grunts, Sobs, Laughs, Startled
Joy	Beautiful, Eternal, Passion, Happiness, Sadness
Sadness	Sorrow, Loneliness, Emptiness, Despair, Joy
Surprise	Invitation, Party, Disappointed, Gifts, Shock
Anticipation	Exhaustion, Horror, Discomfort, Unending, Awakening
Love	Hate, Wonderful, Sweet, Beautiful, Charming
Curiosity	Beliefs, Ignorance, Guilt, Memories, Thrive
Aggressiveness	Unstable, Inferior, Increasing, Dominant, Destructive
Pride	Wealth, Dignity, Wisdom, Courage, Freedom

Table 4: Top similar words to primary emotions and some dyads

5. Experiments and Evaluation

5.1. Evaluation of extracted word-embedding

In order to test the practicality of our extracted word-embedding, we run an experiment on our corpus comparing two approach: One uses manual feature selection and Wordnet-Affect (Strapparava et al., 2004) and another uses the word-embedding for automatic feature extractions.

5.1.1. Feature selection approach

Most research agree that emotion words and phrases are the most obvious clue to identify emotions (Mohammad, 2012a; Strapparava and Mihalcea, 2007). Human have developed language to fit their needs of expressing ideas and feelings. Therefore, when describing our emotion, we tend to use some specific words. By picking up on these words, we have a general idea about the emotional direction of the examined text. In this approach, our first set of features is the basic emotion tendency: to express how an input relates to the 8 basic emotions. Wordnet-Affect is employed to interpret the emotion tendency of each and every words in an utterance. If one emotion exists in one word of the input then the corresponding tendency feature will be set to 1 and 0 otherwise.

However, solely relying on text would cause problem for emotion detection system. We should also consider the effect of negation words and phrases. Simply by putting a negation word, we reverse the emotion state of the text. The sentence: *You are not bad at all!* indicate a strong feeling of approval instead of the usual negative feelings from the word *bad*. Moreover, the context of the input also provide valuable information, especially in conversations. Therefore, we then define the second set of features which includes all similar traits. In the end, we have a list of manual selected features as follow:

1. The sum vector of the current input which suggest the *local tendency*.
2. The sum vector of all the utterances in the lexicon that appear in the conversation which provides the *context of the conversation*.
3. The sum vector of the previous utterance in the conversation which also provides the *context of previous exchange* (of what triggered the current emotion).
4. The *polarity* (negative/ positive) score of the sentence.
5. *Features* such as: length, is_it_a_question, is_it_an_exclamatory_sentence, is_there_negation_word.
6. Colocation features: we mine the ISEAR () dataset for phrases that are often appear in a specific emotional situation. If the input include these phrases, we set the binary flag of the corresponding features to 1.

The structure of the network is shown in figure 2: input layer of manually selected features, two hidden layers, a threshold multi-label output layer.

5.1.2. Word-embedding Network: text to vector

We consider a bag-of-features approach to transform the raw input text into vectors form. Therefore, for a piece of text, its representation is the sum vector of all lexical items inside. Because our goal is to predict the emotional labels for each utterance in a conversation, we also have to vectorize the previous utterance and the entire conversation to capture the contextual information . As a result, the vector representation of an utterance is a 300-dimensional-vector

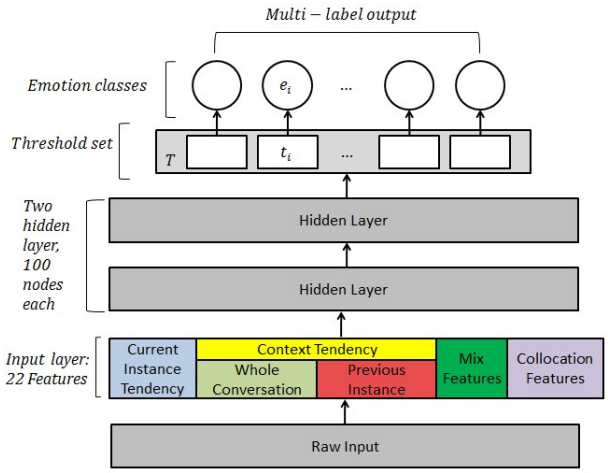


Figure 2: Structure of the manual feature selection network (MFSNet)

concatenated product of the utterance itself and the above-mentioned contextual information. This representation is then fed to the input layer of the neural network in the below figure 3.

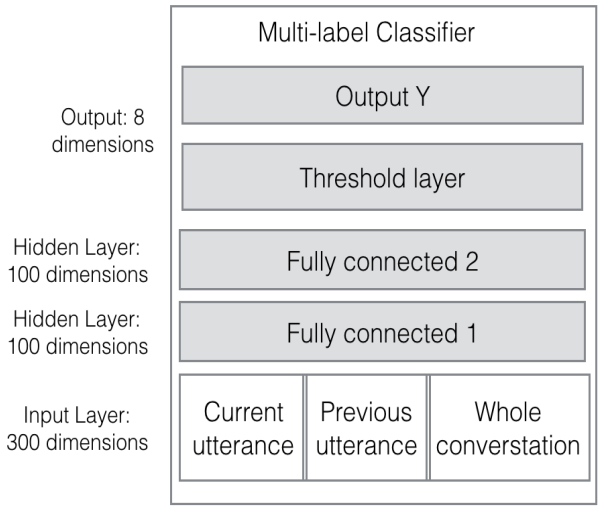


Figure 3: Structure of the word-embedding network

The two networks are both experimented on our annotated corpus. We report the performance of each network and make comparison of our methods to the another method and corpus in the next section.

5.2. Evaluation of the corpus

To evaluate, we use the two previous mentioned networks: manual feature selection network (MFSnet) and word-embeddings network (WENet). We evaluate the result with the gold standard test data using two major measurements in multi-label learning: hamming score (or accuracy in multilabel classification), multilabel F1-score (Table 5). The most important baseline is the average agreement score of the 5 human annotators on our corpus. We also want to compare our corpus to the existing Twitter Emotion Corpus (TEC) (Mohammad, 2012b) for their agreement score and their system's performance. The Twitter Emotion Corpus has tweets with emotion word hashtags. Similar to our work of creating word-embeddings, TEC was used to create the NRC Hashtag Emotion Lexicon (Mohammad, 2012a).

Corpus	Baselines	Hamming score	F1-score
EMTC	**Human annotators**	**43.2**	**62.6**
	WENet	39.1	53.9
	MSFNet	35.1	41.4
TEC	Human annotators	—	43.7
	Binary Classifiers	—	42.2

Table 5: Corpus evaluation

As we can observe, our simple system are performing worse than human annotators by a considerable margin. However, in comparison with other corpora 's Inter-annotators agreement F1-score such as Twitter Emotion Corpus (the score is 43.7), we see the potential of the corpus: It is reliable and the performance of emotion analysis system on it is great. The result is especially significant when our corpus is multilabeled and consists of conversational data which are much more complicated and practical. Both of our networks benefit from the corpus and have good F1-score. While MFSNet is slightly behind Binary Classifiers of TEC, WENet outperforms both. This result suggests that automatic feature extraction using word-embedding is better than manual selected features. We believe that because our embedding are built from the corpus, it has captured the relation between emotional words in the corpus better than general domain lexicon like Wordnet-Affect.

6. Conclusion

In this paper, we present our Emotion Movie Transcript Corpus developed from Imdb quotes dataset. EMTC consists of conversational text extracted from movies and as a result, is close to real-life settings and very practical for emotion analysis tasks. The corpus is partly annotated using our multilabel scheme and the annotators are provided with corresponding movie clips to ensure the reliability of the inter-annotators score of the corpus. We also conduct experiments on two networks: MFSNet that uses manual feature selection and WENet that uses Word-embedding to extract the bag-of-features from the input for supervised learning. The statistics and experimental results show that our extracted word-embedding and the corpus are reliable and even a very simple supervised method like WENet can perform fairly well using only bag-of-features from the embedding.

We would like to investigate the correlation among annotated labels and expand the size of testing data of our corpus using the same annotating scheme in the future. After that, we would focus on building an emotion lexicon from the word-embedding extracted from EMTC.

7. Acknowledgements

This research was supported by JST CREST Grant Number JPMJCR1513, Japan. We deeply appreciate the support of

our colleagues from Computational Linguistics Lab and all the annotators participated in this research for the valuable effort and patience.

8. Bibliographical References

Arimoto, Y. and Okanoya, K. (2016). Comparison of emotional understanding in modality-controlled environments using multimodal online emotional communication corpus. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Choe, W., Chun, H.-S., Noh, J., Lee, S.-D., and Zhang, B.-T. (2013). Estimating multiple evoked emotions from videos. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Collier, G. (2014). *Emotional expression*. Psychology Press.

Dini, L. and Bittar, A. (2016). Emotion analysis on twitter: The hidden challenge. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Hareli, S., Kafetsios, K., and Hess, U. (2015). A cross-cultural study on emotion expression and the learning of social norms. *Frontiers in psychology*, 6:1501.

Hasegawa, T., Kaji, N., Yoshinaga, N., and Toyoda, M. (2013). Predicting and eliciting addressee's emotion in online dialogue. In *ACL (1)*, pages 964–972.

Kruger, J., Epley, N., Parker, J., and Ng, Z.-W. (2005). Egocentrism over e-mail: Can we communicate as well as we think? *Journal of personality and social psychology*, 89(6):925.

Li, M., Long, Y., Qin, L., and Li, W. (2016). Emotion corpus construction based on selection from hashtags. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Liew, J. S. Y., Turtle, H. R., and Liddy, E. D. (2016). Emotweet-28: A fine-grained emotion corpus for sentiment analysis. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Lim, N. (2016). Cultural differences in emotion: differences in emotional arousal level between the east and the west. *Integrative medicine research*, 5(2):105–109.

Mohammad, S. (2012a). Portable features for classifying emotional text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591. Association for Computational Linguistics.

Mohammad, S. M. (2012b). # emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 246–255. Association for Computational Linguistics.

Plutchik, R. (2001). The nature of emotions human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

Scherer, K. R. and Wallbott, H. G. (1994). Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Strapparava, C. and Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics.

Strapparava, C., Valitutti, A., et al. (2004). Wordnet affect: an affective extension of wordnet. Citeseer.

Wang, Z., Lee, S., Li, S., and Zhou, G. (2015). Emotion detection in code-switching texts via bilingual and sentimental information. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 763–768, Beijing, China, July. Association for Computational Linguistics.

Yang, C., Lin, K. H.-Y., and Chen, H.-H. (2007). Emotion classification using web blog corpora. In *Web Intelligence, IEEE/WIC/ACM International Conference on*, pages 275–278. IEEE.