Pakistan Journal of Science, 71 (4 Suppl.): 2019, Page 235-240
ISSN: 2411-0930

HUMAN EMOTION DETECTION AND CLASSIFICATION IN TEXT MINING

S. Ambar*, S. Jan and F. Khan

Department of Computer Software Engineering
University of Engineering and Technology, Mardan, Pakistan
*Corresponding author's E-mail: sadiaamber01@gmail.com

ABSTRACT: This paper presents the enhanced emotion recognition in text and its classification. Emotion detection plays an important role and can be used in wide range of health, business and security applications. The recent research in text mining is mostly based on bipolar approach wherein the emotions are being classified as positive (happy) or negative (angry). In this paper, keyword based approach has been adopted to classify the text into further four emotional categories, i.e., happiness, sadness, anger and other. The proposed approach processes the special linguistic cases like conjunctions and contradictory conjunctions or contrast sentences or its parts to enhance the efficiency of emotion classification. This approach results in twofold advantage: firstly, by considerably enhancing the accuracy of emotion classification and secondly, providing a complete road map to this new area of research. The proposed technique removes emotion stimulus in the text and adjusts the sense of the sentence which improves the accuracy of the classifier.

Keywords: text mining, emotion detection, sentiment analysis, textual emotion classifier

INTRODUCTION

Emotional responses are very difficult to be captured in text, however, it can be gauged computationally by using emotion analysis techniques. There are several techniques to quantify the emotional responses of a person. Emotion is an important aspect of human and is widely studied in social sciences and psychology. Emotion detection in text is a challenging problem because of the absence of voice modulation and facial expressions (Gupta *et al.*, 2017). According to many relevant theories, emotions are triggered by some external event called cause of emotion. The external event is the cause of emotion and not the emotion itself. This paper proposes that removing the cause of emotion in the text can improve the accuracy of emotion recognition in text. We have used a keyword-based approach to classify the text into four emotion categories after handling the contradictory or contrasting conjunction or contrasting sentences and removing the emotional stimulus cause in the text. Emotion analysis is a type of sentiment analysis in which emotions are detected in the text computationally. Emotion analysis refers to the detection of the personal attitude, mood and sentiment towards something. Emotion is an important aspect of human and is widely studied in social sciences and psychology. In addition, emotion detection and its benefits are used in Human-Computer Interaction (HCI), Natural Language Processing (NLP) and machine learning as well. During the last decade, this area of study gained the attention of many researchers, not only in computer science but the emotional aspects play an important role in psychology, communication and healthcare etc. (Chaffar and Inkpen, 2011). Like any other data, emotional analysis can also improve the Customer Relationship Management (CRM).

In contrast to read only, today's read-write web is giving its users the power to interact, post, collaborate and share through social media websites (Cambria *et al.*, 2010). Microblogging is the most popular form of communication. Users can easily express their opinions and emotions through these social media websites. Thus these websites giving rise to huge amount of unstructured data. Analyzing user-generated data for emotion state detection might be useful for applications, such as product recommendation, brand watching and detection of health related issues, etc. (Deyu *et al.*, 2016). The distillation of information from these unstructured textual data is a very complicated task. Many work has been done on emotion recognition on the textual data but the result is far from satisfaction.

Our study focus on to classify the text into three basic emotions Happy, Sad and Anger and all the other emotion into class Other. We are working on linguistic feature and aiming that emotion in the text can be recognize efficiently, if some linguistic features are handled. This paper handle one special linguistic case in NLP, the contradictory or contrasting conjunctions and contrast sentences, sentences which have although, though, but in them. For example, I feel sad for the other contestant but I am happy that I win. We are using keyword base approach aiming that handling the contradictory conjunction can improve the textual emotion classification. Further our study focuses on the detection and removal of cause in the text with emotion stimulus cause, we are aiming that the detection of cause events in the emotional text may increase the emotion classification of text.

MATERIALS AND METHODS

With the information content, text can also communicate the attitudinal information (Aman and Szpakowicz, 2007). Sentiment analysis is typically focusing on classifying the text orientation into positive and negative. The less explore area in the sentiment analysis is the recognition of emotions and its intensity in the text (Aman and Szpakowicz, 2007). Textual emotion detection is a research that is important to analyzing personal emotion in the text (Rachman *et al.*, 2016).

Emotion analysis plays an important role in psychology, communication and healthcare (Chaffar and Inkpen, 2011). The emotion detection and analysis is widely studied in neurosciences, behavior sciences and is important element of human life (Canales and Martínez-Barco, 2014). Online websites are widely used for expressing suicidal thoughts (Desmet and Hoste, 2013). Reference (Desmet and Hoste, 2013) uses emotion analysis in suicidal notes for early risk recognition and prevention of suicides. A large number of email customer may contain complaints about email services. Reference (Gupta *et al.*, 2013) uses emotion analysis to identifies such emotional email (emails with complaints).

Previous studies show that many work has been done to classify the emotion of text, and many approaches has been used to do so. However, the performance of emotion detection is not very satisfactory and many work needs to be done in this context.

Approaches used for textual emotion detection: Reference (Tao, 2004) has used keyword based approach with an enhancement in linguistic information. He divides all the words in the text into emotional functional words and content words, further the emotional function words are divided into emotional keywords, modifiers and metaphor words. Modifiers are the words which influence the emotional keywords such as very, so, too etc., metaphor are the words which have no direct action on the emotion but they do have influence on the emotions for example 'asperity' is more like exaggerating the negative emotion.

Reference (Chaffar and Inkpen, 2011) has used SVM, Bayesian, and decision tree classifiers to categorize text into six basic emotions (anger, disgust, fear, happiness, sadness and surprise). Reference (Chaffar and Inkpen, 2011) applied these algorithms on 3 different datasets.

Reference (Binali *et al.*, 2010) uses a hybrid approach combining both keyword base approach and learning base approach. The output of keyword based approach is used as input for the learning base approach. Ontology is the explicit conceptualization of specification. Ontology can be used for emotion detection (Shivhare and Khethawat, 2012). Reference (Shivhare and Khethawat, 2012) develop emotion ontology using ontology development tool and emotional words hierarchy. Weightage is assigned at each level in the hierarchy. The emotion classifier calculates the weightage of the emotional words.

In general, three approaches are used for emotion detection in text keyword base approach, learning base approach and hybrid approach. These three approaches are the main approaches used. Researcher have used these three approaches with some changes to achieve high accuracy. Keyword base approach:

Predefine keywords are categorized into categories like happy, sad, anger etc. This approach is looking for the keywords in the text and categorize the text based on keywords categories (Tao, 2004).

Learning base Approach: In this approach a train classifier is used to classify text into emotional classes, the keywords in the text is used as features (Binali *et al.*, 2010). Hybrid approach: Hybrid based approach is the combination of both keyword base and learning base approach. A higher accuracy can be achieved with this approach by the use of knowledge rich linguistic information and combination of classifiers (Binali *et al.*, 2010).

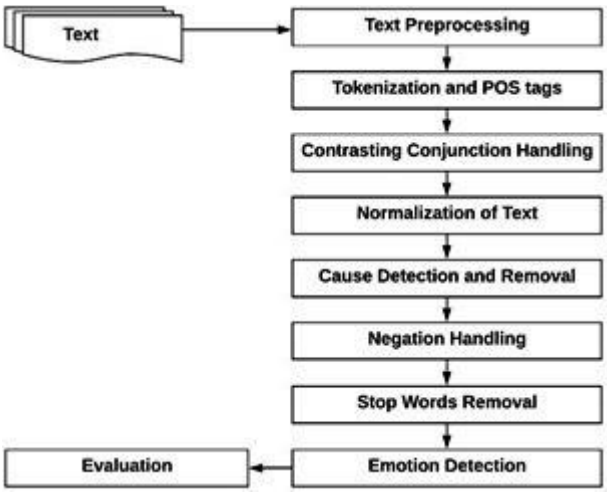

Figure 1. Architecture of the system

Emotion stimulus cause and its detection: Very Little research has been done on the interaction of emotion and the corresponding cause events. The research on the relation between emotion and corresponding cause may play vital role in emotion analysis classification models. Emotion can be invoked by perception of external event and in turn trigger reaction. The detection of the implicit cause of emotion and the actual emotion or reaction in the text can provide rich dimensions to emotion analysis (Lee *et al.*, 2010).

According to many emotional theories, emotion is invoked by external event (Chen *et al.*, 2010). Detecting the cause of the emotion is another NLP issue

and many attempts has been done to detect the cause of the emotion in the text. Reference (Li and Xu, 2014) train the classifier to classify micro- blogging emotion based on emotion cause events.

Reference (Chen *et al.*, 2010) proposed multi-label model which not only detect mutli-clause cause but also capture long distance information to facilitate the detection of the cause of an emotion.

Reference (Song and Meng, 2015) proposes Concept-Level Emotion Cause Model (CECM) for detection of emotion topic and the related cause in short text like microblogging. The CECM uses topic supervised biterm topic model for the detection of 'emotion topic' in event related tweets and context-sensitive topical PageRack is utilize to detect the related cause by detecting meaningful multi-clause expressions.

Proposed Mechanism: This is an experiment base research on text to detect and classify the emotions in text. This research handles contradictory conjunctions or contrast sentences or its parts to improve the efficiency of textual emotion classifier and detect and remove the emotion stimulus cause in text and then emotion of the text is detected and is classified based on lexicon and keywords in the text.

Architecture of the system: Figure 1 shows the steps that are performed to classify the text into emotion categories.

Dataset: The dataset used for this research is the freely available dataset emotion-stimulus data [1], which are made by (Ghazi *et al.*, 2015). The emotion stimulus dataset contain text with emotional tags Happy, Sad, Surprise, Anger, Disgust, fear, shame. The emotion stimulus cause is captured in cause tags e.g. <cause>text</cause>. The above dataset is best to evaluate our proposed mechanism as it has the emotional texts with emotional tags that we are classifying the text into e.g. Happiness, Sadness, Anger and also contain emotional text with tags Surprise, Disgust, Fear and Shame, our system can classify all these tags into class 'Other'. Secondly this dataset also has 676 emotional text with emotion stimulus cause with them and our research suggest that the emotion of text can be detected accurately after the removal of cause. Our proposed system is working with the detection and removal of emotion stimulus cause in the text. This research classifies the emotion of the text in four categories Happy, Sad, Anger, Other. The text with tags Happy, Sad, and Anger in the emotion-stimulus dataset are used for this research. The emotion class of all other emotional text with tags other than Happy, Sad and Anger are considering as 'Other'. Table 1 Shows the detail of the dataset used.

[1] http://www.site.uottawa.ca/~diana/resources/emotion_stimulus_data/

Our proposed mechanism handles the contradictory conjunctions and contrast sentences. Our system proposed that efficiency of textual emotion classifier can be improved by handling such contrast sentences. The above dataset contains 251 contrast sentences or parts of sentences to evaluate our system on.

Table 1. Detail of dataset.

Emotion	Number of text with no Cause	Number of text with Cause
Happy	268	211
Sad	468	107
Anger	276	199
Other	257	159
total	1269	676

Text Preprocessing: Text preprocessing includes the removal of HTML and XML tags, removal of squares and brackets, expansion of words e.g. don't into do not and the case of the text is converted into lower case. We are using beautiful Soup python library for the removal of HTML and XML tags, Contraction library for the expansion of words. A built-in package in python re (regular expression operation) which can be used to work with the Regular Expressions are used for the removal of brackets.

Tokenization: Tokens are the individual words and tokenization in splitting or breaking the text into individual words. In this step the text is breakup into words called tokens, the blank spaces are removed from the text. Our system is using NLTK (natural language toolkit) for tokenization of text.

POS Tagging: Each word in the text is tagged with POS (Part of speech) tags e.g. Noun, Verb, Adjective, Conjunction etc. We are using NLTK to tag each word with POS tags.

Contrasting or Contradictory Conjunctions Handling: This step is the main step in the context of our research. Handling the contradictory conjunctions and contrast sentences can improve the efficiency of the system. This step involves the main algorithm to handle the contradictory conjunctions which will be discussed in section B.

Text Normalization: The text normalization includes the removal of non ASCII words, removal of punctuation marks, and text lemmatization. This system used re and NTLK library for removal of punctuation and lemmatization.

Emotion Cause Detection and Removal: In the context of our research this is also the main step which include algorithm to detect and remove the emotion stimulus

cause in the text. This step will be discussed in detail in section C.

Negation Handling: The sentences that negate the meaning of the emotional keywords is handled in this section. E.g. the sentence "I am not happy", here in this sentence the "happy" word will be classified as Happiness but the emotion of the sentence is "Sadness".

Stop-word removal: Stop words are the most frequent words with less or no meaning like I, am, the etc. All these words are removed in this step. NLTK library is used for the removal of stop word in this system.

Emotion Detection: This step contains the main algorithm to detect the emotion of the text and is explain in section D.

Evaluation: The proposed system is evaluated through information retrieval units, the accuracy, F-score, recall and precision.

Contrasting or Contradictory Conjunction Handling: Some conjunctions like but, although, though connect ideas that contrast e.g. "People say it's dangerous although its amazing country", "My teacher is very nice but a bit strict."

The emotion of the sentences with words like 'although', 'though', 'But' can't be simply detected from key words in the sentence. For example, look at the sentence "I feel exhausted but I am always glad that I am learning new things". Here the emotion class of the sentence before but is Sadness and after but is Happiness, the overall emotion of this sentence is Happiness. Look at the sentence "She was not happy with her boss but definitely she will be happy after knowing her". Here in this sentence the emotion state of sentence before but is Sadness and that of after but is Happiness, the overall emotion of the sentence is Happiness. Just a keyword based approach is not enough to classify the emotion of these sentences. We are splitting the sentence into two sentences before contradictory conjunction and after contradictory conjunction sentence. We are selecting the sentence only before or after contradictory conjunction based on the tense of that sentence. In the above first sentence, the sentence after but will be consider because of 'Present' tense. In the above second sentence the tense before but is Past and that of after but is Future, we have to consider the sentence after but. We are handling the contradictory or contrasting conjunction or contrasting sentences using the following rules.

1. If we have sentence with present tense that will be considered.
2. If we have no present tense sentence, sentence with future tense will be considered.
3. If we have not present, neither future tense sentence then Past tense will be considered.
4. If we have a sentence in which the tense of both before and after contradictory conjunction have the same tense, then sentence after contradictory conjunction will be considered.

Detecting the Tense of Sentences: The algorithm used to detect the tense of the sentence uses POS tags. The sentences which have 'am', 'are', 'is', 'have', 'has' in them are detected as present tense, the sentences which have 'was', 'were', 'had' in them are detected as past tense. If there is text which don't have any word in the above to detect tense of the sentence, then the algorithm is counting the words tags. POS tags 'VBP' (verb present), VBG (verb, gerund/present participle taking), VBZ (verb, 3rd person sing. present takes) are counted as present tense tags. VBD (verb, past tense took), VBN (verb, past participle taken) are counted as past tense tags and MD (modal could, will) are counted as future tense tag. Tense of the text is the one with the greatest number of tags.

Emotion Cause Detection and Removal: Cause of the emotion is some event that trigger the emotion. In some texts like "I am happy that all of your problems just get resolved" the text "all of your problems just get resolved" is the cause of the emotion in the text. The cause has the word 'Problem' which is a Sad word, but the emotion of the text is Happiness. Using the keyword and lexicon based emotion classification the emotion of the above text will be neutral because of both the 'happy' and the 'problem' words in the text, which belong to Happiness and Sadness. Similarly, the sentence like "I am happy taking the tough route". The text "taking the tough route" is the cause of the sentence. In this the emotion of the text is Happiness but because of the word 'Tough' in the cause which is a sad word the emotion of the text will be classifies as neutral. Removing the cause from the text can improve the result of textual emotion classification.

Emotion Cause Detection: The cause of the emotion in text are detected by searching adjective/noun – [Cause key word or VBG verb, gerund/present participle taking]-[anything until it find noun and the words in the detected cause is greater than three or end of the sentence]-noun/end pattern in the text.

The following are the steps of algorithm to detect the cause of emotion.

1. The algorithm is looking for adjective or noun followed by the cause keyword or any verb with POS tag 'VBG'.
2. If pattern in step one found, the algorithm will capture anything as the cause of sentence until it find noun or end of sentence
3. If the algorithm find noun, and the words in the detected cause is greater than or equal to three the captures phrase is the detected cause of the sentence.

4. If the algorithm find noun, and the words in the detected cause is less than three then the algorithm is looking for the next noun or end in the sentence till than all the words are considered as the cause of emotion.

The above pattern found are the detected cause of emotion in the text and is removed from the text. The common cause keywords are being, taking, having, with, that, at, when, of, because, to, for, in, about.

Emotion Detection: We are using a lexicon base and keyword base emotion classification approach. The lexicon we are using are taken from **NRC Word-Emotion Association Lexicon (**Mohammad and Turney, 2013; Mohammad, 2011; Mohammad and Yang, 2011) and **WordNet Effect lexicon.** The keywords in the text are matched with the pre- classified words in the lexicon. The emotional class with the highest number of keywords in the text are considered as the emotion class of that text.

RESULTS AND DISCUSSION

We evaluated the proposed mechanism using widely used information retrieval units the precision, recall, accuracy and F-score. Several tests were applied on the system, test on the proposed system, tests before and after the detection and removal of emotion stimulus cause and test before and after the handling of contradictory sentences to check our proposed work accuracy and F-score. The result of the proposed system is shown in Table 2 below. The Table 2 shows the result of our proposed work, we have get an accuracy of 0.5964 and F- score of 0.5929, based on previous studies this accuracy and F- score shows an improvement from previous emotion detection techniques. Table 3 shows the accuracy and F-score before and after the removal of emotion stimulus cause in the texts.

Table 2. Result of the proposed mechanism.

Dataset Detail	No of Texts	Accuracy	Recall	Precision	F-Score
text with and without cause	1945	0.5964	0.5838	0.6022	0.5929

Table 3 shows the accuracy of the texts with emotion stimulus cause is 0.5636 which improves significantly after the removal of the cause in the text to 0.5784. Similarly, the F-score also improves from 0.5504 to 0.5595 with the removal of cause event from the text. However, the accuracy and F-score of all the corpus data (which include 676 emotional text with emotion stimuli cause and 1269 emotional text without cause event in them) decreases from 0.6077 to 0.5964 and 0.6127 to 0.5929 respectively. The reason in the decrease of accuracy and F-Score of all the data is because that the cause detection algorithm incorrectly detects and remove essential text from sentence which do not have cause with them. The improvement in the result of the "Text with Cause" shows that we can use this approach to improve the efficiency of textual emotion classifiers but more work need to be done to accurately detect the emotion stimulus cause in the texts. Table 4 shows the accuracy and precision before and after the handling of contradictory conjunction and contrasting sentences.

Table 3. Comparism before and after the removal of emotion stimulus cause.

Dataset Detail	No of Texts	Accuracy (before removal of cause)	Accuracy (After removal of cause)	F –Score (before removal of cause)	F-Score (after removal of cause)
Text With Cause	676	0.5636	0.5784	0.5504	0.5595
text with and without cause	1945	0.6077	0.5964	0.6127	0.5929

Table 4. Comparism before and after the handling of contradictory or contrasting conjunctions.

Dataset Detail	Accuracy (before handling of contrasting conjunction)	Accuracy (after handling contrasting conjunction)	F –Score (before handling of contrasting conjunction)	F-Score (after handling of contrasting conjunction)
1945 texts (251 sentences contain contrasting conjunction)	0.5928	0.5964	0.5883	0.5929

Table 4 shows the accuracy before handling of contradictory or contrasting conjunction is 0.5928 and is improve significantly after handling of these Contrasting Sentences. The accuracy after handling of Contradictory or contrasting conjunction or contrasting sentences is 0.5964. Similarly, with the handling of these contrasting sentences the F- score also increase from 0.5883 to 0.5929. The improvement in the result shows that we can improve the performance of emotion detection by handling such cases in NLP.

Conclusion: In this paper we classify the text into four emotional classes Happiness, Sadness, Anger and Other. The keyword base approach is used for classification of emotion in the text. This paper handle one special linguistic case, the sentences that have contradictory conjunction in them. It is concluded that handling the contradictory conjunction or contrasting sentences can improve the accuracy as the accuracy of the classifier increases from 0.5928 to 0.5964 and F-Score from 0.5883 to 0.5929 when tested on the discussed corpus data. This paper suggests that removing the cause of emotion can improve the efficiency of textual emotion classifier. The accuracy of the classifier increases when tested on text with cause from 0.5636 to 0.5784 and F-Score from 0.5504 to 0.5595. However, Cause cannot be detected accurately. Detecting the cause of emotion is a challenging NLP Issue. In future we will work on emotion stimulus cause detection to achieve high accuracy in textual emotion detection.

REFERENCES

Aman, S. and S. Szpakowicz (2007). Identifying expressions of emotion in text. Paper presented at the International Conference on Text, Speech and Dialogue.

Binali, H., C. Wu and V. Potdar (2010). Computational approaches for emotion detection in text. Paper presented at the 4th IEEE International Conference on Digital Ecosystems and Technologies.

Cambria, E., R. Speer, C. Havasi and A. Hussain (2010). Senticnet: A publicly available semantic resource for opinion mining. Paper presented at the 2010 AAAI Fall Symposium Series.

Canales, L. and P. Martínez-Barco (2014). Emotion detection from text: A survey. Paper presented at the Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC).

Chaffar, S. and D. Inkpen (2011). Using a heterogeneous dataset for emotion analysis in text. Paper presented at the Canadian conference on artificial intelligence.

Chen, Y., S.Y.M. Lee, S. Li and C.R. Huang (2010). Emotion cause detection with linguistic constructions. Paper presented at the Proceedings of the 23rd International Conference on Computational Linguistics.

Desmet, B. and V. Hoste (2013). Emotion detection in suicide notes. Expert Systems with Applications, 40(16), 6351-6358.

Deyu, Z., X. Zhang, Y. Zhou, Q. Zhao and X. Geng (2016). Emotion distribution learning from texts. Paper presented at the Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.

Ghazi, D., D. Inkpen and S. Szpakowicz (2015). Detecting emotion stimuli in emotion-bearing sentences. Paper presented at the International Conference on Intelligent Text Processing and Computational Linguistics.

Gupta, N., M. Gilbert and G.D. Fabbrizio (2013). Emotion detection in email customer care. Computational Intelligence, 29(3), 489-505.

Gupta, U., A. Chatterjee, R. Srikanth and P. Agrawal (2017). A sentiment-and-semantics-based approach for emotion detection in textual conversations. arXiv preprint arXiv:1707.06996.

Lee, S.Y.M., Y. Chen and C.R. Huang (2010). A text-driven rule-based system for emotion cause detection. Paper presented at the Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.

Li, W. and H. Xu (2014). Text-based emotion classification using emotion cause extraction. Expert Systems with Applications, 41(4), 1742-1749.

Mohammad, S. (2011). From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. Paper presented at the Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities.

Mohammad, S.M. and P.D. Turney (2013). Crowdsourcing a word–emotion association lexicon. Computational Intelligence, 29(3), 436-465.

Mohammad, S.M. and T.W. Yang (2011). Tracking sentiment in mail: How genders differ on emotional axes. Paper presented at the Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis.

Rachman, F.H., R. Sarno and C. Fatichah (2016). CBE: Corpus-based of emotion for emotion detection in text document. Paper presented at the 2016 3rd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE).

Shivhare, S.N. and S. Khethawat (2012). Emotion detection from text. arXiv preprint arXiv:1205.4944.

Song, S. and Y. Meng (2015). Detecting concept-level emotion cause in microblogging. Paper presented at the Proceedings of the 24th International Conference on World Wide Web.

Tao, J. (2004). Context based emotion detection from text input. Paper presented at the Eighth International Conference on Spoken Language Processing.