International Journal of Engineering & Technology, 7 (2.27) (2018) 27-31

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Emotion detection on cross platform languages

Mohammad Arif [1] *, Mandeep Singh[1], Rajdavinder Boparai[1]

[1]Department of CSE, Chandigarh University, India
**Corresponding author E-mail: official.mohammadarif@gmail.com*

Abstract

Internet has changed the course of our living. It has become the most beneficial antecedent or source of information. Today almost everything is found on internet. Everyday millions of people post their ideas, reviews, stories about the services, products or other persons. The size of data is increasing tremendously. It is very difficult to analyze that amount of data and figure out the emotions or sentiments posed by people. Emotion detection is such a technique where we can judge people's ideas and extract the emotion towards an entity or service. We have used subjective lexicon-based approach to bench the emoticons expressed by the ideas of the people. The data set that we have mainly focused is very cross and noisy. We have used Facebook data in Urdu and Kashmiri language. Both languages are very cross domain. These languages can be written in English alphabet that makes them more challenging to analyses. Our approach resolves the challenge to the maximum possible way. The results shown by our method on this kind of data set are better than any other approach. Our analysis on this type of dataset will help the local businessmen of these areas to grow and flourish. The analysis will give some insights what the local feel about the entity or product so that the manufacturers can design or build it that way and try to enhance its qualities.

Keywords: *Cross Platform Language; Emotion Detection; Social Networks; Subjective Lexicon-Based Approach.*

1. Introduction

Emotion detection is a technique by which user data is analyzed and emotions are derived from it [1]. Emotion detection is the latest and trending research area. With the advancements in the era of technology, huge data gets generated daily. People use social networking sites like twitter and post their stories, ideas, interests, details on them [2]. Four out of six people use social networking sites and almost everyone posts something on them [3]. The significance of data is increasing due to its analytical value but the complexities of handling and processing is also increasing [4]. The data people post on these sites is processed for different analytical applications like generating business insights, political scope etc. for example people on Facebook post their likes and dislikes, this data is processed and insights are generated. Suppose if a user is interested in watching Hollywood movies, he is shown ads related to new release of Hollywood movies. This all is the play of data analytics. Three decades past from today, data had very least significance but with the elevation in technology, a lot has changed [5]. Now we are falling short of storage. In today's era, market shopping has moved to online shopping, libraries have switched to digital libraries and OPAC's, post offices to emails, pen and paper to social networks. With all these advancements in technology, lot of data is getting generated. people can post anything they wish to post and can use any language they are comfortable with. A lot of work has been done on emotion detection but most of that is based on English language. Cross platform languages like Urdu, Kashmiri etc. are not preferred for processing due to their complexities [6]. But it is the need of hour to analyze data based on these languages. We have mentioned these languages (Urdu and Kashmiri) as cross platform languages because at times one single sentence can convey different meanings and emotions.

2. Emotion detection

Emotions are feelings that a person writes or shows towards an entity, service, product or other person [7]. The emotions of a person towards an entity shows the attitude, behavior from that person towards the target entity. Emotions can be likes, dislikes, happy, sad, satisfied, angry, sarcasm, advice etc. emotions are attached with the person and with his real life [8]. Emotions can be used to benchmark a service, product [9]. shopping sites like amazon provide an option for users to give feedback about the product they have purchased. The user can give the feedback in any language. For example, the user can use his mother tongue to post the feedback. Social networking sites like twitter allow its users to post tweets in any language [10]. A user is not restricted to use only English language. Users can tweet in their mother tongue or any regional tongue they are good with. Furthermore, users can do a lot of spelling mistakes while tweeting. Tweets can contain grammatical errors, hashtags, numerals, shortcut words, emoticons etc [11]. All these things may not be beneficial for analytics of data, we have to focus on following points:

2.1. Subjectivity/objectivity

Before performing the analysis of data we need to separate subjective and objective text. Emotions are only expressed by subjective part [12]. Objective part contains no emotions, it contains facts.
Subjective: Mission impossible is the best action movie.
The sentence expresses positive emotion (best)
Objective: Mission impossible was directed by Brain De Palma.

Copyright © 2018 Mohammad Arif et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

2.2. Polarity

The emotions expressed by the subjective part can be divided in three categories (positive, negative and neutral)
Positive: I love my parents.
Negative: I just hate winters.
Neutral: my parents will be returning home by noon.

2.3. Finding target

Target is an entity over which the emotions are detected [13].
Example: Singapore is a beautiful and developed city. It lies near to equator.
Here Singapore is the target. Beautiful and developed are sentiment words. second line of this example is a fact and contains no emotion.

3. Cross domain languages

The dataset we have used for emotion detection is written languages like Urdu and Kashmiri. Both are very different languages. The writing style of these languages is very different from English. English is written from left to right while as these languages are written from right to left. Urdu and Kashmiri have different grammar than that of English. These languages are so versatile that one sentence in English language can be written in many ways with same meaning. The example is shown in (table 1). You can see that one sentence (come here) in English language has been written in five ways with same meaning. That is why we have classified these languages as cross platform languages.

Table 1: Cross Domain Language Example

Kashmiri language	Kashmiri written in English alphabet	Translation
ولس یور	Walsa yuir	Come here
ولآ یور	Wala yuir	Come here
ول حآز یتن	Wal haz yuir	Come here
یور یی	Yuir ye	Come here
یویصآ	Yuir yisaa	Come here

4. Subjective lexicon based approach

Lexicon means an individual unit or token [14]. In this approach a sentence or document is fragmented into individual lexemes or tokens and every word is checked from the pre-defined dictionary of positive and negative words [15]. The overall polarity is calculated by summing the score of words. for example, beautiful, soft, funny, nice are positive words and express positive emotion. While as angry, doubt, envy, offensive are negative words. The dictionary of polarity words can be created manually or can be downloaded from internet. Although emotions can be calculated by using polarity words but there are still some problems that we need to take care of, these problems are as following:

a) Some posts do not contain any sentiment words. we call that as objective text. For example, sun sets in the west.
b) Some posts express emotions without containing any polarity words. Example, my order was placed after two hours (clearly this expresses a negative sentiment i.e. order was delayed).
c) When two or more polarity words are used in combination. Example, the chef of this restaurant cooks terribly sweet food (here terrible expresses a negative emotion and sweet expresses a positive emotion but when both are used together they express super positive emotion).
d) Users use emoticons rather than writing text. Example, thumbs up for like and thumbs down for dislike.
e) Users post in their local language. Example is shown in Table 2.

Table 2: Local Text Example

یہاں بہت سردی ہے	it is very cold here

f) Posts with spelling mistakes. Example, I would luv to visit Singapore (luv misspelt).
g) When users use combination of emoticons and text. In this case we prefer the emotion expressed by the emoticon.
h) Some users express both positive and negative emoticons in a same sentence. In that case we analyze the last part of the sentence. For example, people in Singapore are very good and gentle but it is better to stay indoors after evening.

5. Related work

Sentiment analysis or opinion mining found its significance in early twentieth century when internet became common to everyone [16]. A lot of research has been done in this field. This topic aroused the interest of many researchers. Early studies in this field focus on framework, polarity detection, lexicon creation, feature extraction. Emotion detection created sensation in first text retrieval conference (TREC) [17]. Researchers formulated their task and interpreted how to mathematically incorporate social context and emotions they express. Earlier researchers that sentiments or emotions are segregated by min of a person and they cannot be formulated or benched [18]. But with the emergence of emotion detection, researchers have successfully formulated human emotions into mathematical design. Emotions are extracted and added together. The ideas people share are benched with mathematical and statistical tools and sentiments are derived from them [19]. The hypothetical mind setup of earlier researchers has been implemented now. Twitter dataset was first successfully analyzed for emotion detection and the results were breath taking. Some authors analyzed the posts of engineering students to find the glitches and problems in their educational experiences. They used the tweeter dataset for this analysis [20]. They found that the students had problems like load of subjects, assignments, sleep deficiency. Later on, more approaches were implemented on datasets like Movies, novels, elections etc. Most of the work in emotion detection has been performed on English language [21]. English language is very easy to analyze as there are lots of tools and methods to analyze it. English is an internationally known language and most of the content found on internet is in English [22]. The below mentioned approaches work good with English alphabet. A lot of research has been done using these methods to extract emotions. But the dataset that we have chosen is totally different from English language.
Emotion detection can be done by many other methods also. But we have used subjective lexicon-based approach for our data set because of its format, type and noisy structure. Some of the methods that can be used are as:

5.1. Corpus based approach

This approach is useful for movie reviews, books [23]. The text that should be analyzed by corpus-based approach should contain well-built corpus. Corpus means a word with definite meaning. This model will not yield good results for our dataset as it has a lot of misspellings. For example, I luv Singapore. Luv is misspelt word (love). This method will not show any emotion for this sentence.

5.2. Naïve bayes method

Naïve Bayes method is a machine learning probabilistic approach [24]. Naïve Bayes method falls in the category of supervised learning [25]. This method uses strong dependence features between words and creates a graph where nodes and edges represent random variables and dependencies respectively. This method is expensive to implement and does not yield good results with noisy and unstructured data so, this is not frequently used.

5.3. SVM

Support Vector Machine is a non-probabilistic supervised learning approach where we train a machine by training dataset which labels data and divides them into different classes [26]. When new data is fed into the machine, the algorithm assigns it the labels and divides it into respective classes. Although this method is very effective but the problem is that our dataset contains languages of different platform rather than English so, creating a training dataset is complex process. Furthermore, this method fails to detect sarcasm emotions.

6. Proposed work

We have used subjective lexicon method for our dataset. This method fits our specification and requirement for processing. We downloaded our dataset from respective API's of Facebook and twitter using python. Data was very noisy with lot of misspelling, unwanted fields, grammatical errors, shortcut words etc. we preferred datasets which contained cross platform language like Urdu and Kashmiri. Data formatting was a very crucial step of our analysis. We downloaded polarity words from the internet. There were around 5200 negative words and 2300 positive words. We saved them in two separate text files. We wrote a python program which checked the occurrence of those polarity words in the text and labeled them as negative or positive. The program also calculates the overall emotion expressed by the text or sentence. Some steps after data collection are as following:

6.1. Data formatting

Data formatting is the crucial step of our work. The dataset that we downloaded from Facebook using its API was in csv format, so we converted it into text (txt) format as our python program takes (txt) files as input. Some of the steps in our data formatting are:

6.1.1. N-grams

N grams indicate to successive n terms in the text. One individual term refers to unigram and two terms refer to bigrams and it goes so on to n grams. Sometimes we cannot calculate emotion for unigrams. For example, this work will knock your shoes off.
If shoes off is taken as bigram it will express a positive emotion. If shoes and off are taken as unigrams, it will express a negative emotion.

6.1.2. Parts of speech (POS)

This is one of the important step of data formatting. In a sentence, usually adjectives and adverbs express most of the emotions. So, it is better to tag these words. The highlighted adjectives can directly point out the sentiment or emotion it is expressing and this can save the processing time of the data. For example, weather is too sunny, let us plan a good picnic for this weekend. Sea beach is a good idea. In the above example, the adjectives like sunny, good express the positive sentiment. So, there is no need to further process that sentence.

6.1.3. Stop words

Words like (it, she, he) pronouns are discarded because they provide no or little clue about the emotion. We have to focus on the sentiment, not on who is conveying it or to whom it has been conveyed, so we remove pronouns to make our dataset more refined and noise free.

6.1.4. Stemming

Stemming refers to removing the suffixes and prefixes from the words. We analyse the root word. The algorithm works good for root words as the dictionary we are using contains root words only. So, to make the algorithm working well, we remove the stemming from the words. For example, helping, helped are stemmed to help.

6.1.5. Handling negation

NOT' is a word which can totally invert the emotion. this word totally changes the sentiment to other side. The sentiment words used with NOT are rendered useless. For example, food was not cooked good. Good is a positive word but NOT inverted it to negative.

6.1.6. Handling conjunctions

Some sentences contain a mixture of emotions joined by a conjunction. Conjunctions are used to join two or more sentences. Sometimes the sentiment is expressed by the last part of the sentence. For example, although the place was beautiful but it did not meet my expectations.

6.1.7. Meaning interpretation

The text which was written in mother or regional tongue was translated to its appropriate meaning. Many a times local language was written using English alphabet. For example is shown in Table 3.

Table 3: Meaning Interpretation

Text	Meaning
Aaj Mausam Bohat acha hai	weather is very sunny today
آج ہم نے بہت مزہ آیا	we enjoyed a lot today

6.1.8. Correcting misspelt words

We corrected the misspelt words to their original meaningful words. we simply used Microsoft office and Grammarly software for this purpose. All misspelt words were underlined red and we replaced them with correct spelled words. the words which were underlined blue contained grammatical errors and we corrected them by using Grammarly.

6.1.9. Checking overuse of vowels

Some words in the text contained vowels with exceeding length.
For example, I love this place soooooooo muuuuuuch.
We checked the length of consecutive vowels. The length exceeding 2 or 3 was discarded.
For example, (soooooo to so, muuuuuuch to much).

6.2. Calculating polarity

We wrote a program in python which labels the words as positive or negative and calculates the overall emotion expressed by the sentence. We linked two text files which contain negative and positive words with the program. They act as the dictionary for the program. The program compares every word of the text with the polarity words and assigns the polarity label and score to the words. The total sum of the score expresses the total polarity of the sentence. The work flow of our approach is shown in Figure 1.

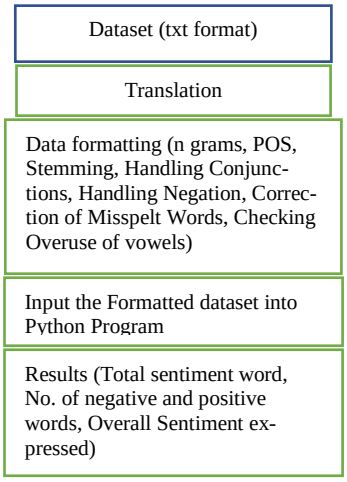


Fig. 1: Method Flow.

7. Results and observations

The results generated by various methods are shown in Table 4. These methods show best results to the dataset that suits their specifications. The downloaded dataset from Facebook API was in csv format with lot of unwanted attributes. We manually removed the extra attributes like (time, genre, location). We converted the dataset into (txt) format. Our dataset contained mainly user posts and comments. The posts that exceeded the length of 200 characters were not taken into consideration. Each post and comment were separated by a separator according to the syntax of our program. The program took one comment in one iteration and calculate the sentiment words, number of positive and negative, and total sentiment. The sentiment calculated was kept on a scale of 10. The value below 5 was treated as negative, value above 5 was treated as positive and value equal to 5 as neutral sentiment. Table 6 shows result generated on three posts, that were posted in Urdu language.

Table 4: Comparison

Method	Task	Dataset	Disadvantages	Results
Corpus based approach	Sentiment Analysis	Novels, movies	Efficiency reduces if dataset is noisy and text has spelling mistakes	Best
Naïve Bayes	Sentiment Classification	Product Reviews	Requires large data set and works on single domain	Good
SVM	Sentiment Classification	Novels, books	Works on single proper domain	Good
Subjective Lexicon	Sentiment Analysis	Noisy, unstructured data, social network data	Depends on dictionary words	Best

Table 5 shows some data formatting done with the data. Each row shows different example of the formatting

Table 5: Data Formatting

Original text	Operation done for modification	Updated text
یہ ہوٹل اس شہر میں بہترین ہے	Urdu to English translation	this hotel is the best in this city
Ye jagah furniture k liye mehshoor hai	Urdu text written in English alphabet	This place is famous for furniture.
It is toooo hot today	Exceeding length of vowel	It is too hot today
I would luv to visit Singapore	Spelling mistakes	I would love to visit Singapore
We can all jump from burj Khalifa without a parachute once in our life	Sarcasm detected	Negative emotion

Table 6: Results

Original text	Translation	Number of sentiment words	Overall Sentiment
آج میرا ےشّم تھا اور میرا پیپر بوہت اچھا ہوا	Today was my exam and I performed well.	1	Positive
مجے جنگلے می رہنا اچھا لگتا ہے مگر وہا جنگلی جانورو کا خطرہ ب ہوتا ہے	I love to stay in forests but there is always a danger of wild animals	3	Negative
آج بوہت گرمی ہے	It is too hot today	1	Negative

8. Applications

The advantages of performing emotion detection on local or regional text will help the local business and politics. Rather than focusing on international trends we should try to let our local people flourish their business. Extracting emotions from local text and generating business insights, political insights will definitely help our local community. Emotion detection helps in business competition. The companies can analyze their user data and act accordingly what users want. This will help the company to grow and compete with other organizations. As we know that business flourishes with the positive feedback from the users. Knowing the customer feedback and sorting out all negative remarks from the users will help the business to regain the trust of customers. Political parties can implement emotion detection before their election campaigns. They an analyze all voter feedback and fulfill their needs to get a heavy vote bank during elections. Law and order department can also take initiative to use emotion detection technique to solve problems of people.

9. Conclusion

This paper tries to remove the barrier between emotion detection and cross platform languages. Most of the work is done on English language or corpus, local or regional languages are not preferred due to data noise, and structure. in this paper we have tried to resolve some of the issues with the cross-platform languages. More work needs to be done to remove all barriers. Although the influence of this paper is limited to certain languages but there are a lot of users on social networks who use these languages. the dataset using these languages cannot be ignored. we can summarize that subjective lexicon-based approach is suited for unstructured data with noise, misspelling and cross domain with different languages. Kashmiri language is a total different dialect and can be best analyzed by lexicon approach rather than any other method. The reviews, comments, blogs, in Kashmiri language can be analyzed and emotions can be detected from them to know what people want and feel about an entity so that the business organization can improve the quality according the need of people.

References

[1] N. Pannala, C. Nawarathna, J. Jayakody, L. Rupasinghe and K. Krishnadeva, "Supervised Learning Based Approach to Aspect Based Sentiment Analysis", *2016 IEEE International Conference on Computer and Information Technology (CIT)*, 2016. https://doi.org/10.1109/CIT.2016.107.

[2] Chenghua Lin, Yulan He, Richard Everson, Member, IEEE, and Stefan Ru¨ger," Weakly Supervised Joint Sentiment-Topic Detection from Text", IEEE trans. on knowledge and data engineering, vol. 24, no. 6, June 2012. https://doi.org/10.1109/TKDE.2011.48.

[3] M. Paredes-Valverde, R. Colomo-Palacios, M. Salas-Zárate and R. Valencia-García, "Sentiment Analysis in Spanish for Improvement of Products and Services: A Deep Learning Approach", Scientific Programming, vol. 2017, pp. 1-6, 2017. https://doi.org/10.1155/2017/1329281.

[4] Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase-level sentiment analysis. In:Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Lin-

guistics, Stroudsburg, PA, USA. https://doi.org/10.3115/1220575.1220619.
[5] Muhammaad Zubair, Aurangzeb Khan, Shakeel Ahmad, Fazal-MasudKundi and Asghar, 2014.A Review of Feature Extraction in Sentiment Analysis. ISSN 2090-4304 Journal of Basic and Applied Scientific Research.
[6] M. Al-Kabi, A. Gigieh, I. Alsmadi, H. Wahsheh, and M. Haidar, "An Opinion Analysis Tool for Colloquial and Standard Arabic," In The fourth International Conference on Information and Communication Systems (ICICS 2013), 2013.
[7] Moreo A, Romero M, Castro JL, Zurita JM. "Lexicon-based commentsoriented news sentiment analyzer system" Expert Syst Appl, 39:9166–80, 2012. https://doi.org/10.1016/j.eswa.2012.02.057.
[8] H Saif, M Fernández, Y He, H Alani (2014),On stopwords, filtering and data sparsity for sentiment analysis of Twitter.
[9] Kang Hanhoon, Yoo Seong Joon, Han Dongil., "Senti-lexicon and improved Naı¨ve Bayes algorithms for sentiment analysis of restaurant reviews", Expert Syst Appl, 39:6000–10, 2012. https://doi.org/10.1016/j.eswa.2011.11.107.
[10] V. M. Pradhan, J. Vala, and P. Balani, "A Survey on Sentiment Analysis Algorithms for Opinion Mining," Int. J. Comput. Appl., vol. 133, no. 9, 2016.
[11] K. Ahmed, N. El Tazi, and A. H. Hossny, "Sentiment Analysis over Social Networks: An Overview," in 2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2015, no. October. https://doi.org/10.1109/SMC.2015.380.
[12] A Semantic web based filtering techniques through web service recommendation", *International Journal of Engineering & Technology*, vol. 7, no. 2-7, p. 41, 2018.
[13] T. V.R. Sai, S. Haaris and S. Sridevi, "Website evaluation using opinion mining", *International Journal of Engineering & Technology*, vol. 7, no. 2-7, p. 51, 2018. https://doi.org/10.14419/ijet.v7i2.7.10257.
[14] "A Method for finding threated web sites through crime data mining and sentiment analysis", *International Journal of Engineering & Technology*, vol. 7, no. 2-7, p. 62, 2018.
[15] User behavior analysis on agriculture mining system", *International Journal of Engineering & Technology*, vol. 7, no. 2-7, p. 37, 2018.
[16] V. Ramya and K. Rao, "Sentiment Analysis of Movie Review using Machine Learning Techniques", vol. &, no. 7, 2018.
[17] Xin Chen,Mihaela Vorvoreanu, and Krishna Madhavan," Mining Social Media Data for Understanding Students' Learning Experiences" *IEEE* trans. on learning technologies, vol. 7, no. 3, JulySeptember 2014. https://doi.org/10.1109/TLT.2013.2296520.
[18] Danushka Bollegala, David Weir, and John Carroll," Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus", *IEEE* trans. on knowledge and data engineering, vol. 25, no. 8, August 2013 https://doi.org/10.1109/TKDE.2012.103.
[19] Xiaohui Yu,Yang Liu, Jimmy Xiangji Huang, and Aijun An,," Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain", *IEEE Trans.On* knowledge and data engineering, vol. 24, no. 4, April 2012 https://doi.org/10.1109/TKDE.2010.269.
[20] G. Salton, "Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer", Addison-Wesley, 1989.
[21] Carlo Strapparava, Rada Mihalcea, "Learning to Identify Emotions in Text", SAC'08 Fortaleza, Brazil 2008 https://doi.org/10.1145/1363686.1364052.
[22] Wiebe, J., "Learning subjective adjectives from corpora", Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000), Austin, Texas, 2000
[23] A. Kathuria and S. Upadhay, "International Journal of Computer Science and Mobile Computing", International Journal of Computer Science and Mobile Computing, vol. 6, no. 4, pp. 17-22, 2017.
[24] M. Gaur and J. Pruthi, "A Survey on Sentiment Analysis and Opinion Mining", international journal of current engineering and technology, vol. 7, no. 2, 2017.
[25] Z. Nanli, Z. Ping, L. Weiguo and C. Meng, "Sentiment analysis: A literature review", 2012 International Symposium on Management of Technology (ISMOT), 2012. https://doi.org/10.1109/ISMOT.2012.6679538.
[26] T. Shivaprasad and J. Shetty, "Sentiment analysis of product reviews: A review", 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), 2017. https://doi.org/10.1109/ICICCT.2017.7975207.