Comment Regeneration and Sentiment Analysis Using Pattern Matching for Facebook

Kamaljot Singh
Department of Computer Science and Engineering
Lovely Professional University, Jalandhar, Punjab, India.
Kamaljotsingh2009@gmail.com

***Abstract* — Sentiment Analysis is the process of analysing the emotional polarity of text. In other words, it determines that whether a piece of writing is depicting a positive or negative sentiment. We proposed to perform sentiment analysis on Facebook comments and present (1) a process to transform noisy comments to analysable piece of writing by pattern matching using numerous dictionaries like slang, regular expression, emotional lookup, good-bad word, scrabble word dictionaries and termed this process as comment regeneration process (CRP), and (2) a technique to analyse the sentiment polarity of the comments using statistical scoring based strategy for comment sentimental analysis (CSA). Incorporating the CRP process, we are able to regenerate the comments by 89% and 84.4% true positives T_P are achieved by statistical scoring strategy based technique for CSA.**

Keywords - Comment Regeneration; Sentimental analysis; polarity analysis; slangs; pattern matching; CSA; CRP

I. INTRODUCTION

Sentiment Analysis is the process of analyzing the emotional polarity of text. In other words, it determines that whether a piece of writing is depicting a positive or negative sentiment. The basic task in comment sentiment analysis (CSA) is classifying the polarity of comment, that whether it is depicting a positive or negative sentiment [1].

The rise of social networking services in daily computing like advertisements, marketing had boosted the interest of researchers towards sentimental analysis. With the proliferation of reviews, rating, recommendations, online opinion and other forms of online expressions, [1]comments has turned into a vital currency for businesses. As the businesses look to automate the process of filtering out the noise, understanding the user opinions, identification of the relevant content and actioning it appropriately, many are now looking to the field of sentiment analysis [2] – [5].

An example of a positive sentiment comment is "*Wee luv it :)*" space and a negative sentiment comment is "*wht a terible movie*". Here, these 2 comments are abstracted from the comments corpus and they suffer from the problem like

- *Regular expression words* – Wee
- *Misspelled words* – terrible
- *Use of emotional symbols* – ":)"
- *Slang* – luv, wht

As, the original comments are very noisy and contains many slang, misspelled, emotional smile's in the text, which must be identified and corrected before any analysis [6]. In this paper, it is proposed to pre-process and to perform sentiment analysis on Facebook comments and present

a) A process to transform noisy comments to analyzable piece of writing by pattern matching using numerous dictionaries like slang, regular expression, emotional lookup, good-bad word, scrabble word dictionaries. This process is termed as comment regeneration process (CRP). The comments are extremely noisy, totally written in informal manner and many slang are used while commenting. This process deals with these noisy comments and tries to reduce such noise.

b) A technique to analyze the sentiment polarity of the comments using statistical scoring based strategy for comment sentimental analysis (CSA). This technique uses statistical scoring based method to classify the comment as positive or negative.

The sentimental analysis can be done in two levels: (1) *Sentence level:* In this, the analysis is carried out by considering the context of word in sentence, (2) *Word level:* In this, the analysis is carried out by considering the positivity and negativity of the word itself. However, for the sentimental analysis in case of comments particularly, we cannot perform the sentence level analysis because of heavy noise and informal nature of comments. So, we used the word level analysis for the proposed work.

This paper is organized as section II discusses about some related works. Problem formulation, comment regeneration and comment sentiment analysis is described in Section III, IV and V. Section VI, deals with experimental results and discussion. Finally, the paper is closed with Conclusion and Future Scope in Section VII followed by acknowledgement and References.

II. RELATED WORK

Ortigosa A. etal [7], had presented a technique for sentiment analysis on messages by extracting the facebook messages. He studied the sentimental changes in messages and implemented an application named *SentBulk*, using same proposed technique. His proposed system had achieved an accuracy of 83.27%.

[1] It is an important activity in social sites, that gives potential to become a discussion forum and it is only one measure of popularity/interest towards post is to which extent readers are inspired to leave comments on document/post.

Kumar P. et.al [8], had automated the process of sentimental analysis on end-user formal reviews. He proposed to filter the irrelevant and unhelpful reviews and quantities thousands of the reviews. His proposed technique generates a summarized report of quantified data.

Usha, M.S. [9], presented a model called combined sentiment topic(CST) model to detect the topic and sentiment simultaneously from text using Gibbs sampling algorithm. This paper formulates this technique as a classification problem to make it more suitable to other domains.

III. Problem Formulation

The two problems of sentiment analysis that are uncovered in this research are (1) comment regeneration and (2) comment sentiment analysis.

The comment regeneration problem is formulated by CRP process that uses *pattern matching* technique to transform the noisy comments to an analyzable piece of writing using slang, regular expression, emotional lookup, good-bad word, scrabble word dictionaries. The task behind this transformation is to match patterns of comments with dictionaries and the identified portion of comment is taken as an accuracy of the process.

Whereas, the other problem i.e.: sentimental analysis on comments is formulated by CSA technique that uses statistical scoring based method to classify the sentiment polarity of the comment [10]. The task here is to identify the words having positive sentiment and negative sentiment with help of below mentioned dictionaries and the training data.

This CRP and CSA requires several dictionaries for regeneration and sentiment polarity classification process. Depending upon the usage of dictionaries, they are categorized as identification type dictionaries and detect-correct pair type dictionaries.

A. *Identification type dictionaries*

These types of dictionaries are used to check the correctness of the words of comment. Each word that appeared in comment is searched in dictionary, if it matches with any word, it is tagged as correct word, otherwise it remained untagged.

1) Scrabble dictionary: The scrabble dictionary is used for the identification of correct words out of the comment. This dictionary consist of 1,78,691 words of English. It was originally developed to play scrabble game. We had used the same dictionary as a part of this work.

2) Ordinary Dictionary: This is a regular dictionary, used to check the correctness of the words of the comment. This dictionary is developed from more than 100 English Novels and online articles.

3) Emotions lookup Dictionary: This is a list of notable and commonly used emotions or textual portrayals of a writer's moods or facial expressions in form of icons. We had collected 45 emotional symbols for this work as ":-)", ":)", ":D", ":3", etc for HAPPY, ">:[", ":-(", ";(", etc for SAD, ":-||", ":@"etc for ANGRY emotions.

4) Dictionary from Training data: This dictionary is developed from the Training data that is provided for training the sentimental polarity classifier. This dictionary contains the word its positivity and it negativity, totalling to one.

5) Google positive and negative word dictionary: Google had developed a positive and a negative word dictionary, we had used that dictionary for identification of polarity of the sentiment.

B. *Detect-correct pair type dictionaries*

1) Slang Dictionary: A type of language consisting of words and phrases that are regarded as very informal, are more common in speech than writing, and are typically restricted to a particular context or group of people. e.g: "10x", "tnx" for "thanks", "tmoro", "tmrw", "tmrz" for "tomorrow" etc. This dictionary contains 6,956 slang words. Out of which 5,433 are common slang, and 1,523 are rejected slang words.

2) Regular Expression Dictionary: This dictionary consist of series of characters and meta-characters also known as wild cards, rather than whole words. This dictionary maps those characters and meta-characters to whole words e.g: "pl+z+" for "Please", "he+l+o+" for "Hello", "me+" for "me" etc. Here "+" is a wild card [11], that is used.

IV. Comment Regeneration Process (CRP)

To regenerate a comment, the raw comment is subjected to be passed through series of sub-processes. Each sub-process is linked with a separate dictionary like slang, scrabble, regular expression, ordinary etc. and follows the pattern matching process to identify and to transform the comment. Figure 1, demonstrates the whole CRP process.

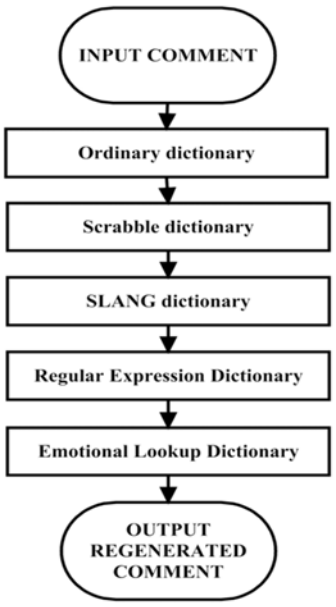

Figure 1: Comment Regeneration Process.

The raw comment is given as an input to this CRP process, which firstly creates the tokens of the comment and then pass those tokens through the dictionaries one by one. The tokens are sent to ordinary and scrabble sub-process, which identifies the correct words by searching the tokens in ordinary and scrabble dictionary, if the token is matched, then that particular token is tagged as identified otherwise it remained untagged.

Then the tokens are forwarded to detect-correct type dictionaries through slang and regular expressions sub-processes. These sub-processes make use of slang and regular expression dictionary to identify natural language words and converts them to standard form as "tnx" to "thanks" or "he+l+o+" to "Hello".

The portion of comment that is tagged or replaced by standard words is compared against original raw comment and the identified portion is taken as an accuracy of the comment regeneration process and is denoted by

$$Accuracy = \frac{Words\_Identified\_or\_Corrected}{Total\_Words} \; x \; 100 \tag{1}$$

where, *Words_Identified_or_Corrected* is the count of tagged or converted words and *Total_Words* is the total number of words in comment.

V. COMMENT SENTIMENTAL ANALYSIS (CSA)

This technique analyses the sentiment polarity of the comments using statistical scoring based strategy and then it classifies the comment as positive sentiment comment or negative sentiment comment. It makes use of various dictionaries like emotional lookup, Google positive and negative word, dictionary from training data dictionaries that are discussed earlier to calculate the positivity and negativity of the comment. Algorithm 1, defines the whole technique for the calculation of positivity and negativity of comment.

Algorithm 1: CSA(Tagged_CORPUS)

This algorithm accepts a hand tagged CORPUS of comments.

Step 1: The main CORPUS is separated into 2 parts, one part is used as training data(larger) and the other part(smaller) is used to test the performance of technique.

Step 2: The training data is processed and a dictionary file is developed out of it. This dictionary contains the sentimental positivity and negativity of all the words that appeared in comments.

Step 3: After processing the training file, these following steps are repeated for every comment.

Step 3.1: The comment is passed through the emotional lookup table. If any emotional symbol is there, the comment is directly classified as positive or negative as per that emotion and the remaining steps are omitted.

Step 3.2: Then, the comment is checked for good or bad words.

Here if, the positive word or good word is identified then the positivity of comment is incremented by 1, and if any negativity is found, the negativity of comment is incremented by 1. If, nothing found then the remaining words are passed to next step.

Step 3.3: At this step the vulgar words has been checked, if the vulgarity is found, the negativity of the comment will be increment by 2,

Step 3.4: After the emotional symbols, good-bad words, vulgarity check of comment, the remaining words are searched in the training dictionary.

Step 4: At last, if the positivity of the comment is higher, the polarity of comment is classified as positive other wise negative.

After regeneration through CRP process, the comment is passed through various sub-process of CSA process as defined in Algorithm 1. The *Step 3* and *Step 4*, is repeated for all the comments.

A. Evaluation Metrics

Results for sentimental analysis are obtained using confusion matrix. The confusion matrix is a tabular visualization of the performance of an algorithm, typically a supervised learning. The instances of predicted class are represented in each column whereas actual class are represented in each row. The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).

The matrix contains 4 fields and number of different metrics are evaluated from them. The fields are True Positive(T_P), True Negative(T_N) also known as Type-II error, False Positive(F_P) also known as Type-I error and False Negative(F_N). The metrics that derive from confusion include:

1) Precision

It is also called positive predictive value(PPV). It is the fraction of classified document that is relevant to user need. It is represented as:

$$Precision = \frac{T_P}{T_P + F_P} \tag{2}$$

2) Recall

It is the fraction of the documents that are successfully retrieved after the execution of query.

$$Recall = \frac{T_P}{T_P + F_N} \tag{3}$$

3) Accuracy

The quality or state of being correct or precise.

$$Accuracy = \frac{T_P+T_N}{T_P+T_N+F_P+F_N} \quad (4)$$

4) F-Measure

F-measure is a measure of test's accuracy.

$$F_1 = 2\,\frac{P.R}{P+R} \quad (5)$$

It considers both the precision *P* and the recall *R* of the test to compute the score.

VI. Experiments

The experiments are performed using JAVA programming language. The crawler uses FQL(Facebook Query language) for crawling the comments from Facebook pages.

A. Experimental Setup

For the Experiments, we had crawled Facebook pages for comments using an extended crawler, that was developed as a part of research in [12]. In total 11,000 comments are crawled and then regenerated using CRP, to reduce noisy data. After Regeneration of the comment data, Whole dataset is hand tagged for the sentimental polarity and final dataset is developed.

The tagged dataset is then broken into 2 parts, in such a way that the training data is of 10,000 comments and out of these 10,000 comments, 4900 comments are of positive polarity, 5100 are of negative polarity. The testing data is of 1,000 words and out of 1,000 words, 500 are of positive polarity and 500 are of negative polarity.

Then out of the training data, the keywords are abstracted and a word based dictionary is developed that contains the word and its positivity and its negativity based on frequency of occurrence in positive comment and in negative comment. The testing data is passed to the system by hiding the polarity of the comment and the accuracy of the technique is measured under confusion matrix.

After applying CSA, the original comments and polarity classified comments are compared to obtain results in confusion matrix.

B. Results and Discussion

After using a number of distinct dictionaries, we are able to identify 89% of the comment and tagged it as identified and 11% still remained as unidentified. However, huge amount of noise has been reduced using dictionaries. For polarity based sentimental analysis, the results are obtained using confusion matrix and are as follows:

Out of 1,000 testing examples (500 positive sentiments and 500 negative sentiments), 844 examples found to be correctly classified and detailed as under:

- Out of the 500 positive sentiment examples, 430 are classified as positive and 70 are classified as negative.
- Out of 500 negative examples, 414 are classified as negative and 86 are classified as positive.

The accuracy of the comment sentiment analysis is achieved to be 84.4% demonstrated in Figure 4.

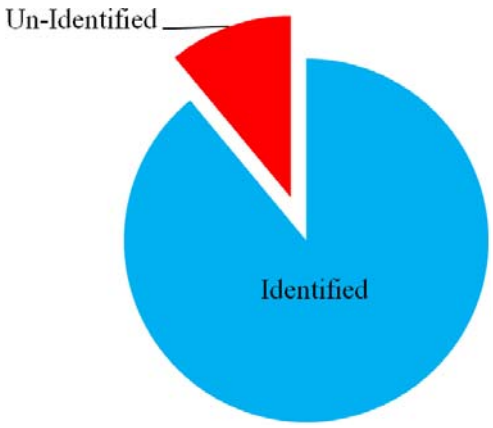

Figure 2. Comment Regeneration Process result.

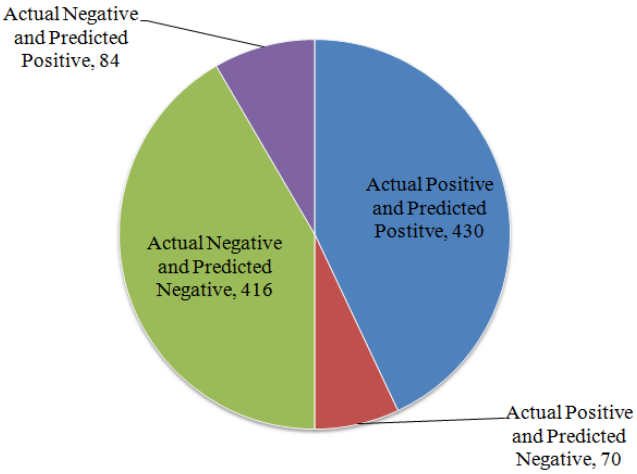

Figure 3. Comment Sentimental Analysis result.

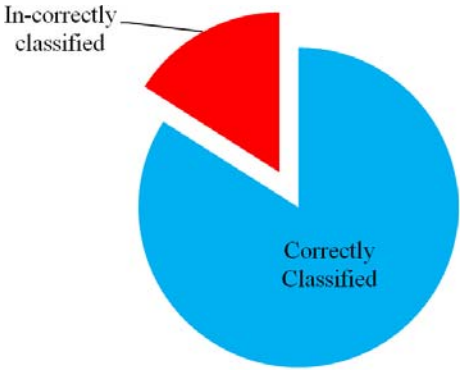

Figure 4. Comment Sentimental analysis.

Therefore, the precision of the system is 83.65% and the Recall of the system comes out to be 86.00% and F1-measure comes out to be 84.80% demonstrating a high precision and stable system.

VII. Conclusion and Future Scope

This is a preliminary work in the field of comment sentiment analysis and we proposed a comment regeneration process (CRP) and comment sentiment analysis (CSA) technique which works fine. By using proposed CRP process, we had regenerated the comment content by 89% and by using proposed CSA technique, we achieved 84.4% true positives T_P. This is also seen that the words in comments can not be processed by the standard context based text processing system because of their high informal nature. This work basically does hard computing based classification over the sentiments, moreover this work can be enhanced by using soft-computing methods for better classification.

Acknowledgment

The authors greatly thanks to all contributors who contributed towards online open source dictionary developments. We would also like to thank Facebook for providing necessary API's for fetching of the comments.

References

[1] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th annual meeting on association for computational linguis- tics*. Association for Computational Linguistics, 2002, pp. 417–424.

[2] A. Aue and M. Gamon, "Customizing sentiment classifiers to new domains: A case study," in *Proceedings of recent advances in natural language processing (RANLP)*, vol. 1, no. 3.1. Citeseer, 2005, pp. 2–1.

[3] G. Mishne and N. S. Glance, "Predicting movie sales from blogger sentiment." in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 2006, pp. 155–158.

[4] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2004, p. 271.

[5] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, vol. 56, no. 4, pp. 82–89, 2013.

[6] M. Gamon, "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis," in *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 2004, p. 841.

[7] A. Ortigosa, J. M. Mart´ın, and R. M. Carro, "Sentiment analysis in Facebook and its application to e-learning," *Computers in Human Behavior*, vol. 31, pp. 527–541, 2014.

[8] P. Kumar Singh, A. Sachdeva, D. Mahajan, N. Pande, and A. Sharma, "An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites," in *Confluence The Next Generation Information Technology Summit (Confluence), 2014 5th International Conference* -, Sept 2014, pp. 329–335.

[9] M. Usha and M. Indra Devi, "Analysis of sentiments using unsupervised learning techniques," in *Information Communication and Embedded Systems (ICICES), 2013 International Conference on*, Feb 2013, pp. 241–245.

[10] I.H.Witten and E.Frank, *DataMining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[11] K. Sierra and B. Bates, *SCJP Sun Certified Programmer for Java 6 Study Guide Exam*, McGrawHill, Ed. McGrawHill Osborne Media, 2008.

[12] K. Singh, R. K. Sandhu, and D. Kumar, "Comment volume prediction using neural networks and decision trees," in *IEEE UKSim-AMSS 17th International Conference on Computer Modelling and Simulation, UKSim2015 (UKSim2015)*, Cambridge, United Kingdom, Mar 2015.

Kamaljot Singh, Assistant professor at Lovely Professional University, INDIA has received his master degree from DAV University in 2015. He had received a Gold medal from NASA, Johnson Space center in 2008 for rover designing. His research interest includes data mining, wireless sensor networks, natural language processing, Nano-magnetic materials and nature inspired computations. He had several publications in reputed journals and in international conferences.