Rating prediction using textual reviews

A NithyaKalyani¹, S Ushasukhanya¹, TYJ Nagamalleswari¹ and S Girija¹
¹ Department of Computer Science Engineering, SRM Institute of Science and Technology, Chennai
{nithyakalyani.a, ushasukhanya.s, nagamalleswari.t, girija.s}@ktr.srmuniv.ac.in

Abstract. Information today is present in the form of opinions. Two & a half quintillion bytes are exchanged today in Internet everyday and a large amount consists of people’s speculation and reflection over an issue. It is the need of the hour to be able to mine this information that is presented to us. Sentimental analysis refers to mining of this raw information to make sense. The discipline of opinion mining has seen a lot of encouragement in the past few years augmented by involvement of social media like Instagram, Facebook, and twitter. The hidden message in this web of information is useful in several fields such as marketing, political polls, product review, forecast market movement, Identifying detractor and promoter. In this endeavor, we introduced sentiment rating system for a particular text or paragraph to determine the opinions polarity. Firstly we resolve the searching problem, tokenization, classification and reliable content identification. Secondly we extract probability for given text or paragraph for both positive & negative sentiment value using naive bayes classifier. At last we use sentiment dictionary (SD), sentiment degree dictionary (SDD) and negation dictionary (ND) for more accuracy. Later we blend all above mentioned factor into given formula to find the rating for the review.

Keywords. Naive Bayes Classifier, SD (Sentiment Dictionary), SDD (Sentiment Degree Dictionary), ND (Negation Dictionary), Social Media, Tokenization.

1. Introduction
We all as a society are subjective creatures and opinions are important to us. Being able to interact with others on that level has many advantages for information systems. This is where sentimental analysis comes into picture. Sentimental analysis refers to process of identifying dominant sentiment in a given piece of text. Information gathered from reviews, comments from various websites like Youtube, Amazon, Flipkart, etc plays very important part of decision making process. Nowadays people tend to buy products with highly reviewed feedbacks, ratings and stars. This clearly shows that a product with more positive sentiment polarity is accepted rather than a product with negative sentiment polarity. Though user’s sentiments are hard to predict but through processed information we can estimate the overall sentiment polarity for the given text by user. The foremost beneficence of our approach are following: a) Here we suggest a system that finds polarity of the sentiment simply by summation of all positive and negative words in the given text, b) We also consider Naive Bayes classifier for finding probability of the given text by user in both forms of polarity that is positive & negative, c) Lastly we suggest sentiment measurement approach based on mined sentiment words and their respective degree of polarity. After blending all the factors we expect to get accurate judgment on the polarity of the text.
2. Related Work

Sentiment analysis is conducted on different layers of text. Such layers are phrase based, sentence based, review based. Review [1],[2],[3],[4] and sentence[5] based layers are operated on whole text to attempt classification based on predefined polarity such as positive, negative, neutral on whole text at once. On the other hand phrase based layer [3] analysis tries to extract the sentiment polarity on the subject discussed in the text. Pang and others [3] discusses a context insensitive evaluative lexical method but is inefficient due to mismatch between base valence of term and author’s usage. Vivek et al [6] discusses about Naive Bayes classifier can be modified to increase and match the accuracy with other complicated models for sentimental analysis using methods like feature selection, n-grams, negation handling.

3. Proposed Work

The main aim of our approach is to find effective sentiment polarity of the given text review by user. In this paper we firstly processed the given input data to remove stop words. The sequences in which the words come are considered by using n-gram concept. In summation to that we described three additional factors for finding sentiment polarity.

![Figure 1. Count of Positive and Negative Words](image)

3.1 Sentiment Identification

Based on [7],[8] HowNet Sentiment Dictionary, we have used SD(Sentiment Dictionary) which contains 8938 words both positive(4363) and negative(4575). The positive sentiment word list contains words with positive polarity such as good, nice, pleasure, happiness etc. The negative sentiment word list contains words with negative polarity such as bad, worst, stink, mistake etc. While deciding the score for given input we count total number of positive words (P) and total number of negative words (N), and then we calculate P – N (RW). Here in figure 1, RW signifies total sentiment polarity of review based on total number of positive and negative words. If the result comes on positive number scale that means review has positive polarity and if the result comes on negative number scale that means review has negative polarity. Zero signifies neutral polarity for review.

3.2 Sentiment Degree

We have used Sentiment Degree Dictionary (SDD) based on the same paper for the words classification on the basis of different levels of degree. SDD contains five different levels. L-1 (52 words) consists of words with highest degree of emotions such as absolute, fully, all, sharply etc. L-2 (48 words) consists of words with higher degree of emotions such as above, how, however, overall
etc. L-3 (12 words) consists words such as more, even, such etc. L-4 (7 words) consists words such as a little, a bit, more or less etc. L-5 (9 words) represent lowest level of emotion degree, consists of words such as less, bit, not very etc. While calculating for the score (figure 2), the words in each clause are considered. We also use Negation Dictionary (ND) which consists of 50 unique features such as no, not, unlike etc. Score is calculated by (-1) ND where ND represents total count of feature in each clause that are found in Negation Dictionary. The degree constant is given in below table:

| Level   | Degree | Words constant(D_w)            |
|---------|--------|--------------------------------|
| Level-1 | 5      | Fully, All                     |
| Level-2 | 4      | Above, Overall                 |
| Level-3 | 2      | More, Even                     |
| Level-4 | 0.5    | A little, A bit                |
| Level-5 | 0.25   | Less, Bit, Not very            |

3.3 Sentiment Probability

Here we use [6] Naive Bayes classifier (figure 3.) which is a probabilistic model, found on Bayes rule with well built independence assumption. Words are classified into two classes positive and negative and are conditionally independent of each other.
Though this classifier does not affect the efficiency and accuracy of text classification, but also makes classification algorithm faster.

3.3.1 Negation Handling. It is a very important factor while determining the sentiment polarity, if not handled accordingly it can distort the final output. For example great refers to positive sentiment but the phrase “Not great enough” is clearly a negative sentiment. Also since we are considering each word as a feature “great” in the given phrase will contribute to only positive polarity. To solve the negation problem we use bi-grams (N-grams) which are taking two consecutive words as a pair and consider it as a single feature or word.

4. System Design
Here in this section we propose a model (figure 4.) on sentiment rating for which we use three different techniques.
Each technique gives polarity of the given text input. Later we fuse all of them to obtain the final result then the result is normalized and the final sentiment rating is obtained. Formula for sentiment calculation:

\[
SC = \frac{NB}{NC} + e^{\left\{ RW^*(-1)ND + Dw \right\}}.
\]

To normalize the score for the sentiments we use:

\[
Ns = \frac{10}{\left\{1 + e^{-\left(Sc\right)\right\}}} - 5.
\]

Table 2. Terms in Sentiment Calculation

| Term  | Description                       |
|-------|-----------------------------------|
| SC    | Sentiment Calculation             |
| NB    | Naive Bayes Probability           |
| NC    | Number of clauses                 |
| Rw    | Total No. of positive words – Total No. of negative words |
| ND    | Negation Factor                   |
| Dw    | Degree of Sentiment               |
| Ns    | Normalize sentiment score         |

Consider a real time example below in figure 5. The words in green font refer to sentiment words. The words in blue font refer to sentiment degree words. The words in bright green font refer to negation words. The words with black font refer to conjunction words.

Figure 5. Real time example

5. Result Analysis

Firstly we calculate number of clauses in the given review. As we can see there are total three clause in the given review, so the value of Nc = 3. In clause 1, “such” is a level-3 sentiment degree word so DW = 2 and “great” is a positive sentiment word so RW = 1, so the total sentiment score of clause 1: 1. (-1)0+2 = 3. In clause 2, “Really” is a level-2 sentiment degree word so DW=4 and “friendly” is a positive sentiment word, “high” is a negative sentiment word therefore RW=1-1=0, Also “not” is a Negation word so ND=1, so the total sentiment score of clause 2: 0.(-1)1+4=4. In clause 3, “tidy”, “delicate”, “tasty” refers to positive sentiment words so RW=3, “really” is a level-2 sentiment
degree word so DW=4, so the total sentiment score of clause 3: 3.(-1)0+4=7. The Naive Bayes probability after training 25,000 reviews. Positive Probability = 0.999873. By summation of all three clauses in the given formula for Sentiment Calculation: 0.999873*(3+4+7)/3 = 4.67. Now to normalize the score we put value of SC in formula to calculate Ns = (10/[1+℮^{-4.67}])-5 = 4.9. Therefore the final rating for given input review is 4.9.

6. Conclusion
In this paper, a mining approach is provided to be used in different recommendation systems. In this we fuse the sentiment polarity with probability to provide a more stable score for the recommendation systems. We used classification of reviews into our predetermined labels before giving the score using bag-of-words approach. This introduction of probability helps in securing the faults which may occur in the dictionary used by the users. Our research paper does not cover the graphical sentiment mining that is using emoticons in the reviews. Now a day’s most of the reviews consist smiley so mining emotions will be very useful for understanding user’s sentiments. Another problem which is not attacked is the problem of sarcasm though the probability of sarcasm in review is less but it’s still possible. Another problem is use of unstructured language which is very prevalent today. Users also tend to mix two or more languages for giving reviews which has to be addressed by the review based recommendation systems.

REFERENCES
[1]. Y. Lin, J. Zhang, X. Wang, and A. Zhou, “An information theoretic approach to sentiment polarity classification”, April 2012 ACM International Conference Proceeding Series. 10.1145/2184305.2184313.
[2]. S. Li et al., “Sentiment classification and polarity shifting”, proceedings of the 23rd International Conference on Computational Linguistics, pages 635-643, Beijing, China – August 23 – 27, 2010.
[3] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques”, Proceedings of the ACL-02 conference on Empirical methods in natural language processing, p.79-86, July 06, 2002.
[4]. Duyu Tang, Bing Qin, and Ting Liu. 2015a, “ Learning semantic representations of users and products for document level sentiment classification”, In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (ACL’15). The Association for Computer Linguistics, 1014–1023.
[5]. T. Nakagawa, K. Inui, and S. Kurohashi, “Dependency tree-based sentiment classification using CRFs with hidden variables”, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, page 786–794. Los Angeles, California, Association for Computational Linguistics, (June 2010).
[6]. Narayan V., Arora L., Bhatia A. (2013) Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model. In: Yin H. et al. (eds) Intelligent Data Engineering and Automated Learning – IDEAL 2013. Lecture Notes in Computer Science, vol 8206. Springer, Berlin, Heidelberg.
[7]. W. Zhang, G. Ding, L. Chen, C. Li, and C. Zhang, “Generating virtual ratings from Chinese reviews to augment online recommendations”, ACM Transactions on Intelligent Systems and Technology (TIST), Volume 4 Issue 1, Article No. 9 January 2013.
[8]. Xiaojingang Lei, Xueming Qian and Guoshuai Zhao “Rating Prediction Based on Social Sentiment From Textual Reviews”, IEEE Transactions on Multimedia, Volume: 18, Issue: 9, Pages 1910 – 1921, September 2016.