Feature Based Sentiment Analysis of Product Reviews using Modified PMI-IR method

Sanjay Kalamdhad[1], Shivendra Dubey[2], Mukesh Dixit[3]

[1] *M. Tech Research Scholar, Department of Computer Science REC Bhopal India*
[2] *Research Guide, Department of Computer Science REC Bhopal India*
[3] *Head, Department of Computer Science REC Bhopal India*

Abstract— *In online product reviews users discuss about products and its features. A product may have hundreds or thousands of reviews, consumers share their experience about products and comments about products characteristics. These product reviews may have positive or negative sentiments. A positive sentiment contains good opinion about product and its features similarly a negative sentiment tells drawbacks and problems of product and its features. Feature may be part of the product or its characteristics. In this paper we use modified PMI-IR method for analyzing the sentiments in online product reviews about the various features of products. We download the product reviews from internet using the web crawler and stored it in inverted index format. Using the parts-of speech tagging, extract the two-word opinion phrases and calculates the semantic orientation by measuring the mutual information between each phrases and positivity and negativity. Summary of sentiments of each feature is presented based on average semantic orientation value. Summarization shows the sentiment classification of features of products.*

Keywords: *Sentiment analysis, web crawler, semantic orientation, PMI-IR, summarization*

I. Introduction

Now day's online shopping websites is more popular and convenient medium for selling and buying products for both manufacturers and consumers. Consumers buy almost everything which is available online. Consumers also comment about products they purchase by posting reviews of products. These reviews contain positive or negative sentiments about products. Reviews also discuss about specialty of products i.e. features of products. Many reviews are long and take time for reading; some of them are not related about products. Large collection of reviews and ratings are unable to present detailed information about features of products. It becomes hard for people to find generalized opinion about the particular product features.

Sentiment analysis or opinion mining capture the attention by researchers in last few years which analyze sentiments expressed in written text in English or any other languages. Generally sentiment analysis is categorized in three types; documents based, sentences based and feature based or aspect based. Feature/aspect based sentiment analysis detects polarity of sentiments of features.

Two approaches of sentiment classification used to determine orientation of reviews. One is machine learning based methods [9, 10] and second is semantic oriented methods.

To collect the online product reviews, we first build the web crawler; it is crawled into web pages of shopping site and collects the rating and reviews of products posted by consumers. Ratings and reviews are stored in Inverted Index format so these can be searched easily. Inverted index is special data structures that stores documents as TF-IDF (term freq. – indexed doc. freq.).

Turney [1] presented a semantic oriented based mining method which uses point-wise mutual information and information retrieval (PMI-IR) method for sentiment classification which uses mutual information and statistical data collected by IR. Mutual information described as amount of information about two word-phrases when we observe the positive and negativity. When we talk about something positive or negative, how frequently these opinion phrases co-occur with positive or negative words. It is an unsupervised approach of sentiment classification that detects the polarity on document level.

Turney's algorithm use parts-of-speech tagger to extract the two-word phrases of adjectives, adverbs and some other combination from product reviews. Next is, calculate semantic orientation of each phrase using PMI-IR algorithm. A numerical value is assigned to each phrase which shows the association with positive reference word *('excellent')* and negative reference word (*'poor'*). The association or co-occurrence of phrases with 'excellent' and 'poor' is positive (e.g. "great mobile") or negative (e.g. "bad quality"). Now determine the average semantic orientation that decides the reviews are recommended or not recommended. If the average is positive then review will be considered as useful for product otherwise it will consider as not useful. The numerical value of average semantic orientation also describes the strength of positivity and negativity of the review. The semantic orientation of phrase is calculated using PMI and IR. Mutual information is calculated between each phrase and *'excellent'* and is subtracted from the mutual information of each phrase and *'poor'*. Mutual information is amount of information of the presence of two word phrases when we observe *'excellent* and *'poor'*. When talking about positivity and negativity, frequent occurrence of sentiment phrases along with reference words is mutual information. So phrases that co-occur with *'excellent'* are more likely positive and terms that tend to co-occur with *'poor'* are more likely negative. Turney's algorithm uses 410 reviews from shopping site Epinions collected from four different domains: reviews for travel destinations, banks, automobiles and movies, these are not written by professionals but posted by consumers of Epinoins and express their opinions about products.

II. Sentiment Analysis

For sentiment analyses following tasks are performed on reviews.

- **Web crawling:-** Download ratings and reviews from shopping website using the web crawler and stored in indexing files.
- **Review Sentences: -** Find those sentences where product features are mentioned and parse these review sentences and assign tags to each word using parts-of-speech tagger (NLP technique)
- **POS Tagging: -** Based on certain patterns mentioned in Table I extract two-word phrases from review sentences.
- **Semantic Orientation: -** Calculate the Semantic Orientation of each phrase, PMI-IR algorithm takes *'excellent'* and *'poor'* as reference word.

Semantic Orientation of phrase is calculated as,

$$SO\ (phrase) = PMI\ (phrase, excellent) - PMI\ (phrase, poor) \qquad (1)$$

Here point-wise mutual information (PMI) between word1 and word2 is defined as,

$$\mathrm{PMI}(word1, word2) = \log_2\left[\frac{\mathrm{p}(word1\ \&\ word2)}{\mathrm{p}(word1)\ .\ \mathrm{p}(word2)}\right] \qquad (2)$$

PMI value is calculated by passing queries to search engine and noting the number of results found (number of matching documents). Here p (*word1 & word2)* is the mutual information that *word1* and *word2* occurs with each other. If they occur alone, the mutual information is given by the product of p (*word1*) and p (*word2*). The ratio between p (*word1 & word2)* and p (*word1*) p (*word2*) measures dependency of *word1* over *word2*. The log value of ratio is called PMI value of *word1* that determines correlation with *word2*. PMI value of *word1* is positive when *word1* tend to combine with *word2* and negative when *word2* not available with *word1*. For example hit (*query*) is the number of results (matching documents) returned from the online search engine, for the given query SO (*phrase)* is calculated from the equations (1) and (2) as follows:

$$\mathrm{SO}(phrase) = \log_2\left[\frac{\mathrm{hits}(phrase\ \mathrm{NEAR}\ \text{"excellent"})\ \mathrm{hits}(\text{"poor"})}{\mathrm{hits}(phrase\ \mathrm{NEAR}\ \text{"poor"})\ \mathrm{hits}(\text{"excellent"})}\right] \qquad (3)$$

The NEAR operator constraints search to documents that contains *phrase* and *excellent (or poor)* within a given window size.

A. Phrase extraction: In English language each word is categorized in tags or syntax using parts-of-speech tagger [14]. Few patterns are described in Table I. Following these patterns, two-word phrases extracted from review sentences that contains features of products. These two-word phrases contain adverbs, adjectives, nouns and verbs that show subjectivity and characteristics of products. The patterns that extract two-word phrases are adopted from Turney's study.

Table I .Two-word Phrase Patterns

S.No.	First Word	Second Word
1.	JJ	NN
2.	RB	JJ
3.	JJ	JJ
4.	NN	JJ
5.	RB	VB

The SO of each phrase determines number of positive and negative phrases. If SO is greater than threshold value phrase consider as positive phrase otherwise phrase is negative. Similarly average SO is calculated by adding SO of all phrases, it is positive if greater than threshold value otherwise average SO is considered as negative. Threshold value in turney's study is zero. If average semantic orientation is greater than threshold value, review has positive opinion otherwise negative opinion. Numerical value of average SO determines the strength of positivity and negativity.

III. Related Work

Turney's [2001] first work that uses numerical data calculated by querying online search engine identify synonyms [6] of words. It is a simple unsupervised algorithm called PMI-IR that evaluates statistical resemblance of synonyms of words. Using this PMI-IR algorithm Turney classify online product reviews by extracting the two-word phrases and estimating the semantic orientation of phrases. Product reviews used in his study extracted from four different domains automobiles, banks, movies and travel destinations. Of these 410 reviews 170 are *not recommended* and remaining 240 are *recommended*. The average accuracy of classification algorithm for all four domains is 74%, ranging from 84% for automobile reviews to 66% for movie reviews.

In 1997 Hatzivassiloglou and McKeown [2] proposed a supervised algorithm that predicts the semantic orientation of adjectives but it is designed only for isolated adjectives rather than two-word phrases that contains of patterns of adjectives, adverbs and nouns.

YE Qiang, LI Yijun, ZHANG Yiwen [3] worked on Chinese product reviews. Their study was based on book and cell phone reviews written in Chinese language. They extract two-word phrases from Chinese reviews and calculate the semantic orientation. The orientation of reviews is decided by different threshold values.

Similar work has been done by ZHANG Zi-qiong, LI YI-jun, YE Qiang and LAW Rob [4] in 2008. They use an unsupervised PMI-IR method for sentiment classification of Chinese product reviews. Instead of using number of hits of query they use snippets returned from Google. For example, to calculate PMI value of phrases issue a query and crawl returned snippets.

Sentiment classification of blog contents are determined by calculating semantic orientation. Depending on contexts blogs have different sentiments like joy, angry etc. Xuiting Duan, Tingtin HE, Le SONG [5] study blog content and classify contents as joy, angry, fear, sad using semantic orientation method.

M. Hu and B. Liu [7] and Won Young Kim, Joon Suk Ryu, Kyu II Kim, Ung Mo Kim [13] study the customer reviews for mining opinions and generate the summary of opinions.

X. Ding, B. Liu and PS. Yu [8] suggests a holistic lexicon based approach of feature-based sentiment classification that identify different features of products and detects opinions about features.

Qingliang Miao, Qiudan Li, Ruwei Dai [14] study strategy for finding product features from reviews and classifying opinions about features.

All these methods applied on individual reviews and predict the sentiments from individual review and present overall summary. In our study, instead of estimating opinions of individual reviews we calculate the semantic orientation of all reviews of products and extract overall opinion about features of products and show the amount of positivity and negativity about feature of products. In this paper, we study the sentiment classification of most common features of product that expressed in reviews by consumers. Experimental results show that the method is more viable than recent methods.

IV. Proposed Methodology

A. Web Crawler: - First of all, we develop a web crawler to collect the ratings and reviews of five different online brands of mobiles, tablets and laptops and stored in Inverted Index format in a special file format called indexing files. Contents are stored and arranged in memory as TFIDF, so it is possible search by terms present in files; reason of using indexed files as they reduce memory uses by using natural language techniques like removal of stop-words and using Stemming algorithm.

B. Feature Selection: - To identify features of products from the reviews we use word tokenization and parts-of-speech tagging techniques, where words are converted into tokens. Features like camera, price etc. are nouns and are frequently used in reviews. So we choose most common features for our product categories mobiles, tablets and laptops, these are *camera*, *price*, *battery* and *processor*.

C. POS Tagging: - It’s an important task in NLP and sentiment analysis. Every word in English grammar is nouns, adverbs, adjective etc. POS tagging is a NLP technique of assigning tags to words of review sentences. This technique identifies single nouns and groups of tags of certain patterns mentioned in Tab.1 by using the Chunking and parsing the tree. Given a sentence *“Picture quality of camera is excellent”* generates a tree of POS tags.

For all four features, we find those sentences in which these features are mentioned. These reviews sentences have the meaningful opinion about feature of product.

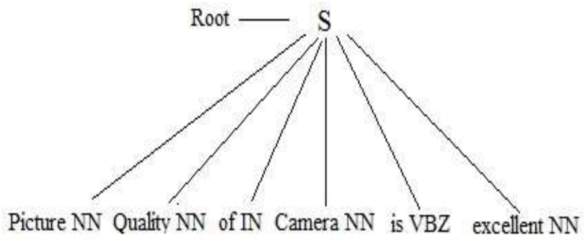


Figure 1.POS Tree Structure

Algorithm for phrase extraction:

This algorithm extract phrases from reviews sentences using following patterns and store in text files.

```
1. define structure for phrase extraction,
        grammar = r””” (chunk:
            {<JJ.*><NN.?>+}{<RB.*><JJ.?>+}
          {<JJ.*><JJ.*>+}{<NN.*><JJ.?>+}
        {<RB.*><VB.?>+}”””
2. parse the grammar using regular expression
3. for each rev_sent of feature f1:
            parse and generate phrase tree
4. for subtree in phrase tree:
        #sub[0][0]and sub[0][1] are left and right
         node of tree
        chunk_list = []
        #add left, right node to list
        chunk_list.append = sub[0][0] + ‘ ’+ sub[0][1]
        #add pos reference words with phrase
        chunkpos1 = sub [0][0] + ‘ ’+ sub[0][1]+’excellent’
        chunkpos2= sub[0][0] + ‘ ’+ sub[0][1]+’great’
        chunkpos3 = sub[0][0] + ‘ ’+ sub[0][1]+’good’
                ………          ………     ……..
        chunkpos18 = sub[0][0] + ‘ ’+ sub[0][1]+’best’
        #add neg reference words with phrase
        chunkneg1 = sub[0][0] + ‘ ’+ sub[0][1]+’poor’
        chunkneg2= sub[0][0] + ‘ ’+ sub[0][1]+’worst’
        chunkneg3 = sub[0][0] + ‘ ’+ sub[0][1]+’bad’
                ………          ………     ……..
        chunkneg18 = sub[0][0] + ‘ ’+ sub[0][1]+’sad’
5. store phrases in text file.
```

D. Modified PMI-IR method: For the calculation of PMI values and semantic orientation we propose two modifications in PMI-IR method. In our study, we use 18 positive and negative reference words. We use 18 positive and negative words instead of *‘excellent’* and *‘poor’*. This improves the efficiency of algorithm by adding more positivity and negativity with phrases.

Positive reference words =
{‘excellent’, ’good’, ’fantastic’, ’best’, ’super’, ‘great’,’amazing’,’awesome’,’stunning’,’beautiful’,’nice’,’worth’,’decent’,’brilliant’,’average’,’extraordinary’,’powerful’,’fine’ }

Negative reference words =
{‘poor’, ’bad’, ’worst’, ’wrong’, ’problem’, ’defective’, ’damage’,’terrible’,’sucks’,’heating’,’heavy’,’pathetic’,’ridiculous’,’regret’,’fault’,’sad’,’annoying’,’useless’,’awful’}

PMI values of phrases are calculated from equation (2), for this we do not use any web search engine, instead we develop a reviews search engine that has more than 20,000 online product reviews. By passing queries to reviews search engine, we estimate number of results (hits) for phrases (i.e. ‘*phrase’ AND pos/neg.ref words*). In equation (2) now, *word2* represents 18 positive and negative words. Unlike other online search engine, Review search engine contains only products reviews which talks about only products and its features nothing else. This improves reliability.

Using equation mentioned below, Semantic Orientation of each phrase is calculated by subtracting PMI values of phrase with positive reference words and PMI values of phrase with negative reference words.

$$\textbf{SO } (\textit{phrase}) = \textbf{PMI } (\textit{phrase, positive words}) - \textbf{PMI } (\textit{phrase, negative words})$$

SO is consider as positive if its value for each phrase exceeds threshold value and negative if it is less than threshold value. Average of SO of all phrases calculated, it is the semantic orientation of the feature. It is believed that positivity in reviews is always greater than negativity. According to Maite Taboada's SO-CAL program [11] positivity in reviews sentences is 1.5 times greater than negativity. In our study we set threshold value as 2.

PMI Algorithm:

```
Procedure pmi_ senti_words(phrase_list)
    for each phr in phrase_list:
        hit_list = []    # empty list
        hit_list += phr
        if hit_list empty:
            'No phrase found'
            exit()
        else
            pmi_calculation(phr)
        end if
    end for
PMI values of reference words
pos = [p1,p2,p3,……p18]
neg = [n1,n2,n3,.....n18]
n = no. of reviews
pos_list = pos(i) / n
neg_list = neg(i) / n
Procedure pmi_calculation(phrase)
    # smoothing operator to avoid division by zero
    hit_list = hit_list[float(i) + 1.0]
    lrg_no = length(hit_list)
    n = no. of reviews
    hit_list = hit_list[float(i) / n + lrg_no]
return hit_list
```

Orientation_calculation *(feature)*

```
Calculate pmi values of phrases with reference words
        senti_words(ref_phrase_words)
Calculate pmi values of two-word phrases
        senti_words(phrase_words)
SO = pmi_senti_words(ref_phrase_words) –
                pmi_senti_words(phrase_words)
define threshold value
if SO >= threshold value:
        'SO is positive'
else if SO < threshold value
        'SO is negative'
Calculate average SO
```

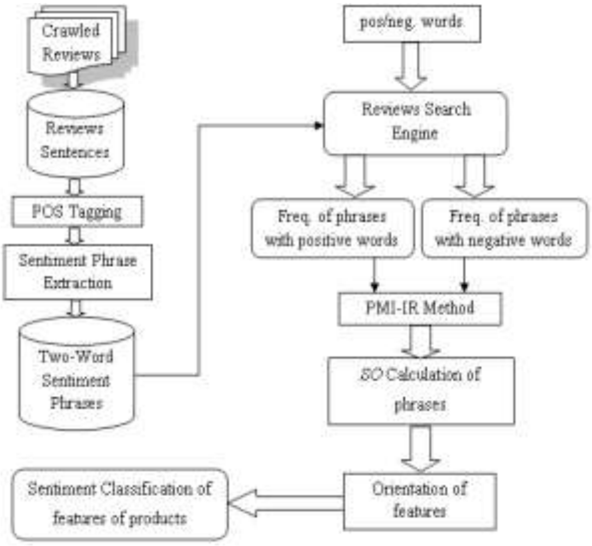

Crawled Reviews
Reviews Sentences
POS Tagging
Sentiment Phrase Extraction
Two-Word Sentiment Phrases
pos/neg. words
Reviews Search Engine
Freq. of phrases with positive words
Freq. of phrases with negative words
PMI-IR Method
SO Calculation of phrases
Orientation of features
Sentiment Classification of features of products

Fig. 1 shows Architecture of overall system

Semantic Orientation of all phrases finds the number of positive and negative phrases; this determines the amount of positivity and negativity for the feature. Averaging the semantic orientation of all phrases shows the strength of opinion of features of products. The orientation of average SO is positive if it is higher than threshold value and if average SO is lower than threshold value. Bigger the average SO depicts higher amount of positivity, similarly smaller average SO depicts higher amount of negativity. This work finds the opinions of overall reviews from our database; it does not focus on the opinion of individual review.

We calculate semantic orientation for all four features and present the summary of all features in percentages, as the amount of positivity and negativity along with strengths. For example features of mobile brand shows summary as;

Summary.......

Product:	*Apple iPhone 6*
Feature:	*camera*
Positivity:	61.62 %
Negativity:	38.46 %
Strength:	5.6261

Product: *Apple iPhone 6*

Feature: *battery*
Positivity: 57.89 %
Negativity: 42.11 %
Strength: 6.4440

V. Experiments and Results

To classify the polarity of features of products, we first download 60 reviews of each product categorized in mobiles, tablets and laptops. Each category has five products, this makes total of 900 product reviews. All these reviews and ratings are downloaded from online shopping website www.flipkart.com using web crawler. In the process of phrase extraction pattern total of 3059 phrases are extracted among them 820 phrases are of feature *camera*, 978 of *battery*, 305 of *processor* and 956 of *price*. Semantic Orientation of all phrases calculated for each feature and based on SO of positive and negative phrases, percentages calculate for each feature.

Table II shows extracted phrases and corresponding semantic orientation. Here positive phrases are more than negative phrases, so positivity for this feature is higher than negativity, also average semantic orientation shows the strength of positivity.

Product Category: *Tablets*
Product name: *Lenovo A7-30*
Feature: *price*

TABLE II

Extracted Phrases	SO
"good product"	18.78066
"lower price"	10.51247
"whole pipeline"	-10.83723
"headphone overall"	-4.912420
"good tab."	5.890250
"good quality."	17.07945
"much happy"	10.00198
"reasonable price"	19.26637
"not bad"	11.05807
"good purchase"	11.38700
"same price"	14.68596
"network easy"	-10.83723
"great price"	20.31622
"also want"	-2.134810
"really awesome"	17.71352
"3g/wi-fi/calling tab"	-11.8372 3
"fantastic price"	2.469010
Average SO	6.9760

Results: Positivity: 70.59 %
Negativity: 29.41 %
Strength: 6.9765

Table III shows extracted phrases and corresponding semantic orientation. Here negative phrases are more than positive phrases, so negativity for this feature is higher than negativity, also average semantic orientation shows the strength of negativity.

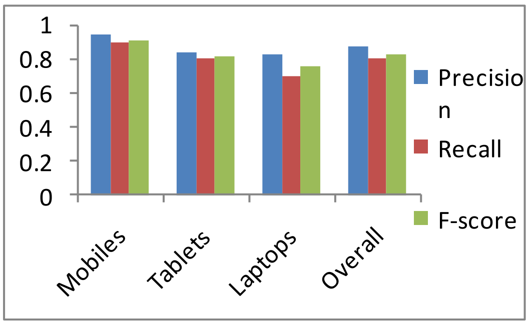
Product Category: *Laptops*
Product Name: *Apple MacBook Pro*
Feature: *processor*

TABLE III

Extracted Phrases	SO
"powerful machine"	2.66808
"3rd gen"	-12.83723
"upgradable upto"	-6.83723
"hrs ....standby"	-2.87144
"even tough"	8.43419
"bright display"	5.71719
"free software"	-1.91242
"wouldnot prefer"	-6.83723
"subwoofer....it surrounds"	-6.34566
"big hall"	-2.49738
"free multi"	-6.83723
"big disadvantage"	-2.28648
"never go"	6.94597
"newer model"	0.08758
"pathetic onboard"	-9.83723
"dual core"	9.30028
Average SO	-1.6325

Results: Positivity: 31.25 %
Negativity: 68.75 %
Strength: -1.6325

Performances of such systems are measured using the standard evaluation parameters like precision, recall and F-score.



Precision (*p*) is ratio of relevant outputs to the retrieved outputs; recall (*r*) is ratio of relevant outputs to total outputs. F-score (*f*) is measured as *2p.r/ (p+r)*. Following table shows all parameters for all three categories of products and also the performance of overall system.

The overall performance of our study is comparable to the results of semantic oriented based approach in the earlier studies for product reviews analysis, which ranges from 70 % to 85 %.

Table 4 .Performance measurements of system

	Precision	*recall*	*f-score*
Mobiles	94.73 %	90.00 %	91.95 %
Tablets	84.21%	80.00%	81.95 %
Laptops	82.35 %	70.00 %	75.52 %
Overall System	87.27 %	80.00 %	83.35 %

VI. Conclusions

This paper presents an unsupervised modified PMI-IR method of classifying opinions of features of online products as positive or negative. Results also show the strength of positivity and negativity of each feature and summary of all features. PMI-IR method is simple and it is not compulsory to use corpora sets to train inputs. Our proposed method uses web crawler to store online reviews, POS tagging which is one of natural language technique, we also develop reviews search engine which has the dataset of 20,000 product reviews. Sentiment classification of features of products is useful for shopping websites where it is possible to give more detailed information about the product from consumer's point of view. Showing user opinions about special features of products is great beneficiary to both online retailer and buyer of products.

For further improvement, we can increase the database of our reviews search engine; bigger the search database will increase the reliability of the system. Phrase extraction patterns are crucial to implement as there is possibility of useless phrases, we expect more specific opinion oriented phrases could be identified from reviews for improving performance.

References

[1] Turney, P. 2002 Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification Reviews *ACL'02*.

[2] Hatzivassiloglou, V and McKeown, K 1997. Predicting the Semantic Orientation of Adjectives, *In proc of 35th ACL/8th EACL.*

[3] Q. Ye, Y. Li, Z. Yiewn, 2005. Semantic Oriented Sentiment Classification For Chinese Products Reviews: An Experimental Study on Books and Cell Phone Reviews.

[4] Z. Zi-qiong, LI Yi-jun, YE Qiang, LAW Rob, International Conference 2008, Sentiment Classification for Chinese Product Review Using an Unsupervised Internet-based Method.

[5] X. DUAN, T. HE, Le SONG, Research on Sentiment Classification of Blog Based on PMI-IR.

[6] Turney P. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL.

[7] M. Hu and B. Liu, Mining and summarizing customer reviews, *KDD'04* 2004.

[8] X. Ding, B. Liu and PS. Yu, 2008 International conference, A Holistic Lexicon Based Approach to Opinion Mining.

[9] Vapnik V. N. The Nature of Statistical Learning Theory, New York: Springer 1998.

[10] Fei, Z. C., Liu J., Wu G.F., Sentiment Classification using Phrase Patterns. In: Proceeding of the 4th International Conference on Computer and Information Technology (CIT'04). Wuhan, China: IEEE, 2004: 1-6.

[11] Maite Taboada's SO-CAL program, Lexicon-based methods for sentiment analysis M Taboada, J Brooke, M Tofiloski, K Voll

[12] NLProcessor - *Text analysis toolkits 2000.* https://www.infogistics.com/textanalysis.html

[13] Won Young Kim, Joon Suk Ryu, Kyu Il Kim, Ung Mo Kim, A Method for Opinion Mining of Product Reviews using Association Rules.

[14] Santorini, B. 1995. *Part-of-Speech Tagging guideline for the Penn Treebank Project* (3rd revision, 2nd Printing), Technical Report, Department of Computer and Information Science, University of Pennsylvania.

[15] A.-M. Popescu, O. Etzioni. Extracting product features and opinions from reviews[C]//Proc. of Conf. on Empirical Methods in Natural Language Processing, EMNLP'05, 2005: 339-346.

[16] M. Gamon, A. Aue. Automatic identification of sentiment vocabulary: Exploiting low association with known sentiment terms[C]//Proc. of the ACL-05 Workshop on Feature Engineering for Machine Learning in Natural Language Processing, 2005:57-64.

[17] Church. K. W. and Hanks, P. 1990, Word Association Norms, Mutual Information and Lexicography.

[18] T. Mullen, N. Collier. Incorporating topic information into sentiment analysis models [C]//Proc. of the ACL 2004 on Interactive poster and demonstration sessions, 2004.

[19] Turney and Littman 2003, Measuring praise and criticism: Inference of semantic orientation from association.

[20] V. Ng, S. Dasgupta and S. M. Niaz Arifin, Examining the Role of Linguistic Knowledge Source in Automatic Identification and Classification of Reviews. *ACL'06*, 2006.

[21] S. Kim and E. Hovy Determine the Sentiment of Opinions.

[22] B. Pang, L. Lee, and S. Vaithyanathan Thumbs up? Sentiment Classification Using Machine Learning Techniques *EMNLP'2002*.