Identification of Co-occurrence Frequencies Based on Sentiment Analysis with Spreading Activation Algorithm

Meda SivaKrishna[1], K Yasudha[2], Vanitha kakollu[3]

[1,2,3]*GITAM (Deemed to be University), Visakhapatnam*

Abstract

Rapid changing in society, online consumer reviews plays a vital role to assist purchase-decision making which has become increasingly popular. It provides a considerable source of consumer reviews, but one can hardly read all reviews to obtain a fair evaluation of a product or serviace. Users of the online shopping site Amazon are encouraged to post reviews of the products that they purchase. Little attempt is made by Amazon to restrict or limit the content of these reviews. The number of reviews for different products varies, but the reviews provide accessible and plentiful data for relatively easy analysis for a range of applications. This paper seeks to apply and extend the current work in the field of natural language processing and sentiment analysis to data retrieved from e-commerce. A text processing framework that can summarize reviews, would therefore be desirable. A sub-task to be performed by such a framework would be to find the general aspect categories addressed in review sentences, for which this paper presents two methods. Among the most existing approaches, the first method presents an unsupervised method that applies association rule mining on co-occurrence frequency data obtained from a corpus to find these aspect categories. While not on par with state-of-the-art supervised methods, the proposed unsupervised method performs better than several simple baselines such as supervised and unsupervised methods.

Index Terms: *Sentiment Analysis, Co-occurrence Data, Notional Words, Stop Words, Stem Words, Lemmas, Consumer Reviews, Spreading Activation.*

1. Introduction

The process of identifying and classify reviews expressed in a piece of text, especially in order to determine whether the consumer's attitude towards a particular topic, product etc., is positive, negative, or neutral. In quintessence, it is the process of determining the behind a series of words, used to gain an understanding of the attitudes, opinions and emotions expressed within an online[13]. The number of reviews a user gives a product is used as training data to perform supervised machine learning. For instance, a corpus contains 15,000 product review from 12 products serves as the dataset of case study[9]. Highest selling and reviewed products on the site are the primary focus of the experiments, but useful features of them that aid in accurate classification are compared to those most useful in classification of other category products. One of the most important forms of text based communication are

product and service reviews, Posted on the web by customers[6]. Retail companies such as Amazon, flipkart and paytm mall have frequent reviews of the products they sell, which provide a superfluity information, and sites like yelp offer detailed consumer reviews of local multi cuisine restaurants, hotels, and other businesses aspects[11]. Research has shown these reviews are considered more valuable for consumers than market-generated information and editorial recommendations and are increasingly used in purchase decision-making[2].

The statistics that can be obtained from product and service reviews is not only beneficial to consumers, but also to companies. Knowing what has been posted on the Web can help companies improve their products or services[1]. Nevertheless, to effectively handle the large amount of information available in these reviews, a framework for the automated summarization of reviews is required. An important task for such a framework would be to recognize the topics (i.e., characteristics of the product or service) people write about[7]. These topics can be fine-grained, in the case of aspect-level sentiment analysis, or more universal in the case of aspect categories.

2. Related work

Major aspect categories are used for detecting unstated fine-grained aspects. These might be used for detail categories as well. An initial work on implicit aspect detection is proposed in this paper to use semantic association analysis based on Point Mutual Information (PMI) to a pair off outcomes x and y belonging from discrete random variables or single notional words.

$$PMI(x, y) = \log \frac{P(x, y)}{P(x)P(y)}$$

In the above formula, it specifies x as the random variable that models the occurrence of a word, and y the occurrence of a class[8]. For a given word x for a given classy, PMI is used to decide if a feature is informative or not, and you can do feature selection on that basis. Furthermore, this work aims to provide encapsulate either positive and negative feedback about products, laws or policies by mining reviews, discussions, forums etc. This approach diligently scans every line of data, and generates a forceful summary of every review (categorized by aspects) along with various graphical visualizations. It proposes rule based hybrid approach[3]. It finds sequential patterns to obtain both explicit and implicit aspects. “Aspect-based analysis it focuses on extraction of aspects from customer reviews and ranking these aspects as positive or negative”.

It aims is to automate the process of gathering online end user reviews for any given product or service and analyzing those reviews in terms of the sentiments expressed about specific features [10]. In e-commerce websites, clients typically provoke comments, which

incorporate those properties of the product, those mentalities of the vendor, express conveyance majority of the data following purchasing the results[4]. The majority of the data gives a critical reference to the point when others purchase results in the website. On assumption analysis and finer-grained idea mining approach concentrates for the resulting features. Regrettably, there were no quantitative experimental results reported in their work, but intuitively the use of statistical semantic association analysis should allow for certain opinion words such as “large,” to estimate the associated aspect (“size”). In an approach is suggested that simultaneously and iteratively clusters product aspects and opinion words. Aspects words with high similarity are clustered together, and aspects/opinion words from different clusters are dissimilar[12]. The similarity between two aspects words is measured by fusing both homogeneous similarity between the aspects words (Text information), calculated by traditional approach, and similarity by their respective heterogeneous relationships they have with the thoughts/aspects (link information). Based on the product aspect categories and opinion set of word groups, a sentiment association set between the two groups is then constructed by identify-ing the strongest *n* sentiment links. This approach, however, only considered adjectives as opinion words which are not able to cover every opinion, yet the approach was capable of finding hidden links between product aspects and adjectives. Unfortunately, there were no quantitative experimental results reported, specifically for implicit aspect identification.

3. Proposed System

In the existing system, a user goes through various websites, reads through the reviews and then determines whether to buy the product or not. In this paper we proposed a system to determine whether the consumer's attitude towards a particular topic, product etc., is positive, negative, or neutral. In quintessence, it is the process of determining the behind a series of words, used to gain an understanding of the attitudes, opinions and emotions expressed within an online[5]. The number of reviews a user gives a product is used as training data to performed with Spreading Activation Algorithm (SAA) using supervised machine learning. The only requirements are that the user should know the specific name of the product he wishes to buy and only the reviews which are in English can be processed.

The reviews for various products can be found on various E-commerce websites like Flipkart, Amazon, EBay, Paytm mall etc. Our primary step is to extract reviews from these sites for further processing. In the training phase, reviews from Amazon have been given to the machine along with customer feedback as input. Then after consider the input and classify the words and group them with respective manner either positive, negative and neutral which is depends upon reviews. For example, transactional data can be mined taking the grocery store for frequent patterns and association rules. For instance, rule could be {milk, coffee powder} -> {sugar}, this rule would imply that if someone purchased milk and coffee powder together, then they are also likely to purchase sugar. Subsequnetially, several algorithms were developed to use association rule mining for classification. During the training stage, the general

notation to label and then mine only those rules that that have a class label on the right hand side.

Association rule mining was used in unsupervised scenarios to discover interesting patterns. These rules are representatives of the corresponding classes, and essentially are the classifiers. When you get a new (unlabeled) instance, you would score it against all of the rules and the class that scored the highest becomes the predicted class label for this new instance. For example, the fantastic category has the seed set {fantastic, tremendous, impressive}. This measure gives an idea of how frequent an item set is in all the transactions. Consider itemset1 = {milk} and itemset2 = {sugar}. There will be far more transactions containing milk than those containing sugar. So as you rightly guessed, itemset1 will generally have a higher support than itemset2. Now consider itemset1 = {milk, curd} and itemset2= {milk, sugar}. Many transactions will have both bread and butter on the cart but bread and shampoo? Not so much. So in this case, itemset1 will generally have a higher support than itemset2. Statistically, support is the fraction of the total number of transactions in which the item set occurs.

$$\text{Support}(\{x\}\text{->}\{y\}) = \frac{\text{Transactions containing both x and y}}{\text{Total number of transactions}}$$

Spreading Activation Algorithm: Here we consider two methods supervised and unsupervised Initially, to identify for each of the given categories $c \in C$ a set of grain words G_c containing the category word and any synonyms of that word. To find the determination of co-occurrence of graph like both vertices and edges The weight of each edge $(i, j) \in E$ is represented by $W_{i,j}$ and denotes the conditional probability that notional word i co-occurs with notional word j in a sentence after it, given that j is present in that sentence. This formula is shown as follows:

$$w_{i,j} = \frac{x_{i,j}}{N_j} \qquad (1)$$

where $X_{i,j}$ is the co-occurrence frequency of words i and j (word i after word j) and N_j is the frequency of word j.

Implementation and Methodology:

The SAA propose in the current accomplishments differs from the original algorithm in a different of ways. First, it considers the two most likely senses for each word group and iteratively disambiguate the word group with the high weighted difference between the similarity of both senses to the corpus text, rather than the word group with the greatest similarity for its best sense. However, this should yield better results than the original

algorithm, as it allows to consider the best separation of the senses of the to-be-disambiguated terms for that we can aspect category detection execution, picking the most similar sense might not be the best option if the similarity difference with respect to the next best sense is small.

First consider the input for SAA as category c, vertices V, seed vertices S_c, weight matrix W, decay factor δ, firing threshold τ_c. Then the following are the steps for SAA

Algorithm 1:

$A_{c,s} \leftarrow 1$
end
foreach $i \in V \setminus S_c$ do
$A_{c,i} \leftarrow 0$
end
$F \leftarrow S_c$
$M \leftarrow S_c$
while $M = \emptyset$ do
For each $i \in M$ do
For each $j \in V$ do
$A_{c,j} \leftarrow \min \{A_{c,j} + A_{c,i} \cdot W_{i,j} \cdot \delta, 1\}$ end
end
$M \leftarrow \emptyset$
foreach $i \in V \setminus F$ do
if $A_{c,i} > \tau_c$ then
add i to F
add i to M
end
end
end

Algorithm 2:

for sentence s ∈ test data do
for aspect category a ∈ A do
customer value = 0
for word w ∈ s do
if O(w) > 0 then
customer value= value + (w, a)/O(w)
end if
end for
sentence = s/ length(s)
if score > threshold value then
Assign aspect category a to s
end if
end for
if s has no assigned aspect categories then
Assign ‘anecdotes/miscellaneous’ to s
end if
end for

Second, we can Determine Weight Matrix W, then all unique categories are identified, storing them in each category set C. In addition, the co-occurrence

frequencies of all lemmas and dependency forms are stored in vector Y, while the co-occurrence frequencies of all dependency form/lemma-category combinations, are counted and stored in matrix X, respectively.

Results Analysis

We analyze the results obtained from the algorithms the analysis is done using two methods, they are Supervised and Unsupervised in this evaluation process of finding similarity differences of two statements (exclusive reviews) like Sequential process of the value 30.8 which is similar to both odd/even and random process. The Following are the results obtained and Output generated.

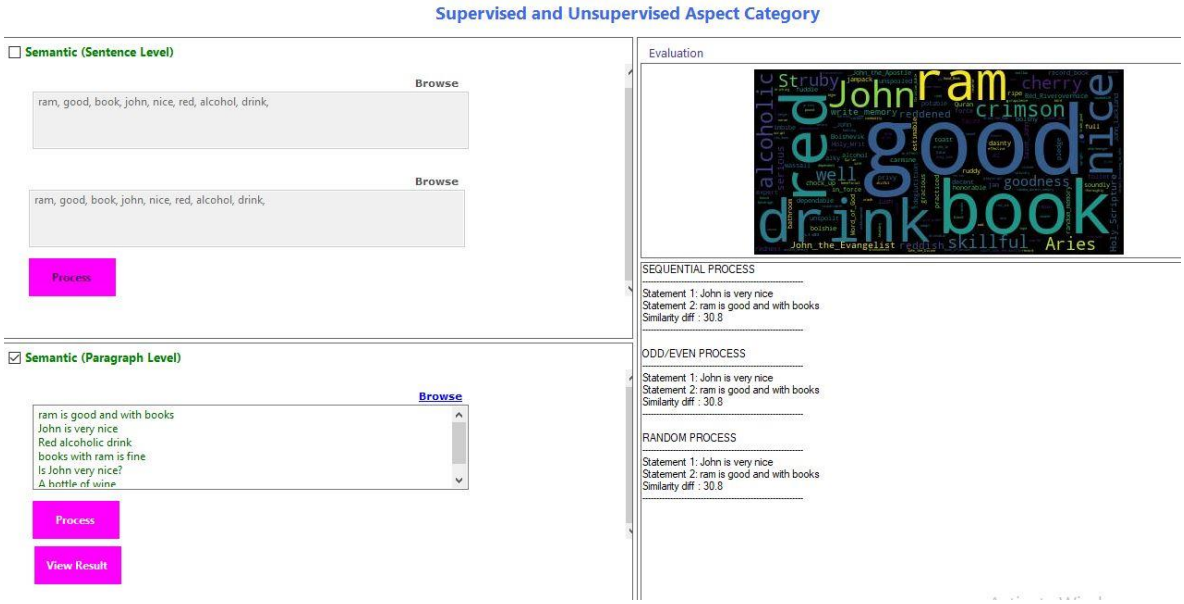

Figure 1: Identifying Similarity Differences

The above figure illustrates the three different process have more refined view of different similarity words expressed in the customer feedback about the products in ecommerce, we should move to the sentence level. This level of sentiment analysis filters out those sentences which contain no differences in their reviews and which is determines whether the customer feedback of the product is positive, negative and neutral.

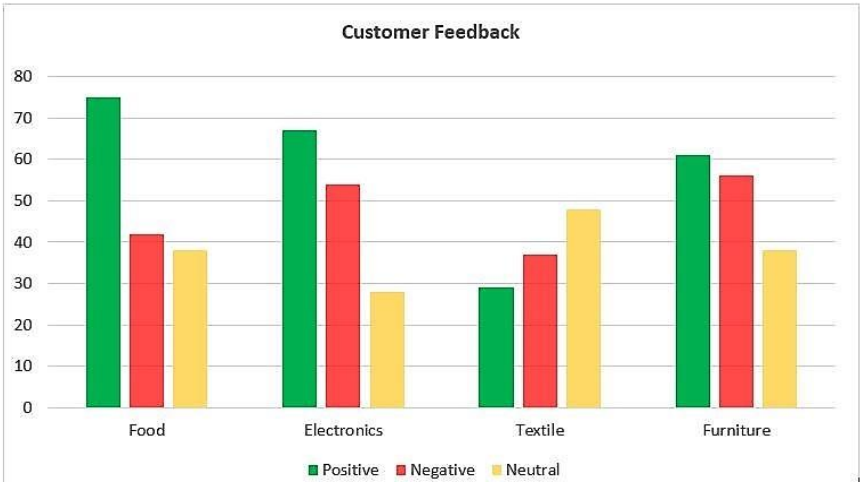

Figure 2: Customer Feedback

The above bar graph demonstrates method uses a list of words and category of expressions used to express people's subjective feelings and sentiment or opinions. It not only uses certain words, but also phrases and idioms. In the other types of sentiment analysis, we have seen what positive and negative words are. Let us take an example: *"food is better than remaining categories."* This sentence does not express an opinion that any of the three types customer reviews positive, negative and neutral. Therefore, these types of sentences/documents are furthered analyzed using text-based approach.

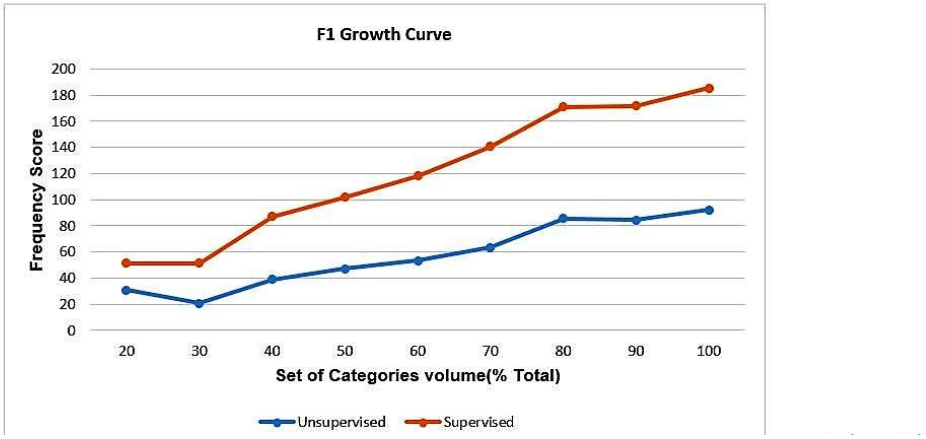

Figure 3: F1-Growth Curve

Whereas line graph establishes identifying F1-growth curve means frequency dependency are shown for different sizes of the training set, using a stratified sampling technique where the distribution of the sets remains similar to the innovative data set. Each data point in the figure epitomizes an incremental intensification of different feedback sentences in labeled data, for the supervised method, and unlabeled data for the unsupervised method. The supervised method continuously seem like to outperform the unsupervised method, even though larger exercise sizes for the unsupervised method seem to accomplish on equivalence with

the supervised method for which very small volumes of labeled data are available which is very accurate value.

Conclusion

In this paper mainly, we proposed a system which uses supervised and unsupervised machine learning algorithm that is Spreading Activation Algorithm. This system mainly focuses on only text based classification. Product reviews can be classified based on the seed words they transmit. This will decrease the overhead of traversing through various sites. This results in every word having an activation value for each category that represents how likely it is to imply that category. While other approaches need labeled training data to operate, this method works unsupervised. If the maximum conditional probability is higher than the associated, trained, threshold, the category is assigned to that sentence. Evaluating this approach on test set shows a high F1-score of 84.67% with collaborative methods.

References

[1] Thumbs up? Sentiment Classification using Machine Learning Techniques. BoPang and LillianLee, Shivakumar Vaithyanathan [IBM, Cornell University].

[2] Pang, B., Lee, L. and Vaithyanathan, "S. Thumbs up? sentiment classification using machine learning techniques," Proceedings of the ACL-02 conference on Empirical methods in natural language Processing-Volume 10 (pp. 79-86). Association for Computational Linguistics, July 2002

[3] Emotions in product reviews – Empirics and models. David Garcia, Frank Schweitzer. Chair of Systems Design, ETH Zurich.

[4] Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis. Geetika Gautam and Divakar Yadav [Jaypee Institute of Information Technology].

[5] Ashok, Meghana, et al, "A personalized recommender system using Machine Learning based Sentiment Analysis over social data," Electrical, Electronics and Computer Science (SCEECS), 2016 IEEE Students' Conference on. IEEE, 2016.

[6] Kim, S.M. and Hovy, E," Determining the sentiment of opinions," In Proceedings of the 20th international conference on Computational Linguistics (p. 1367). Association for Computational Linguistics, August 2004

[7] M. T. Adjei, S. M. Noble, and C. H. Noble, "The influence of C2C communications in online brand communities on customer purchase behavior," J. Acad. Marketing Sci., vol. 38, no. 5, pp. 634–653, 2010.

[8] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Found. Trends Inf. Retrieval, vol. 2, nos. 1–2, pp. 1–135, 2008.

[9] C.-L. Liu, W.-H. Hsaio, C.-H. Lee, G.-C. Lu, and E. Jou, "Movie rating and review summarization in mobile environment," IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 42, no. 3, pp. 397–407, May 2012.

[10] M. Pontiki et al., "SemEval-2014 Task 4: Aspect based sentiment analysis," in Proc. 8th Int. Workshop Semantic Eval. (SemEval), Dublin, Ireland, 2014, pp. 27–35.

[11] S. Kiritchenko, X. Zhu, C. Cherry, and S. M. Mohammad, "NRCCananda-2014: Detecting aspects and sentiment in customer reviews," in Proc. 8th Int. Workshop Semantic Eval. (SemEval), Dublin, Ireland, 2014, pp. 437–442.

[12] T. Brychcin, M. Konkol, and J. Steinberger, "UWB: Machine learning approach to aspect-based sentiment analysis," in Proc. 8th Int. Workshop Semantic Eval. (SemEval), Dublin, Ireland, 2014, pp. 817–822.

[13] C. R. C. Brun, D. N. Popa, and C. Roux, "XRCE: Hybrid classification for aspect-based sentiment analysis," in Proc. 8th Int. Workshop Semantic Eval. (SemEval), Dublin, Ireland, 2014, pp. 838–842.