Opinion Mining for Extraction of Opinion Word and Opinion Target Using Modified Word Alignment Joint Aspect and Sentiment Model

C[1], P.Sengottuvelan[2]
[1]Associate Prof. in Computer Science, Department of Computer Science, Muthuragam Govt. Arts College, Vellore -2
[2]Associate Prof. in Computer Science, Department of Computer Science, PG Extn.Center Periyar University Dharmapuri
[1]anette1967cs@gmail.com, [2]sengottuvelan@gmail.com

ABSTRACT Data mining is a technique of collecting, searching through, and analyzing a huge amount of data stored in a database, as to determine patterns or relationships. Sequences of challenges have emerged in data mining and in that one of the major challenges is opinion mining. Opinion mining is the field of study which works with people appraisals, opinions, sentiments and emotion towards the entities such as products, services. The main aim is to gather the opinion regarding the products from the online review websites. A real online review from different domains is selected as the evaluation datasets. This method is compared with several state-of-the-art methods on opinion target/word extraction. It is assumed that all nouns/noun phrases in sentences are opinion target candidates and all adjectives/verbs are considered as potential opinion words, which are generally adopted by earlier methods.

Keywords – Data Mining, Modified Word Alignment, Opinion Targets, Joint Aspect and Sentimental model.

1. INTRODUCTION

Data mining is the extraction of hidden predictive information from huge databases, is a commanding new technology with immense potential in order to help companies focus on the mainly chief information in their data warehouses.

Data mining tools predicts future trends and behaviours by allowing businesses to make hands-on knowledge-driven decisions. The prospective analysis is obtained by data mining move beyond the analysis of earlier period events provided by retrospective tools typical of decision support systems.

Data mining tools can respond business questions that usually were more time consuming to solve. They search the databases for hidden patterns thereby finding predictive information that experts may fail to notice because it lies outside their expectations.

Mining opinion words and opinion targets from online reviews are chief tasks for fine-grained opinion mining, the main component of which involves detecting opinion relations between words with the quick development of Web 2.0, a massive number of product reviews are springing up on the Web. From these reviews, consumers can obtain first-hand assessments of product in sequence and direct management of their purchase actions.

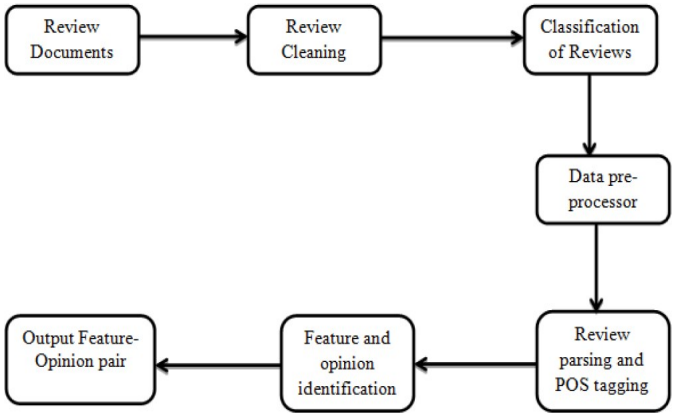

Figure 1. Architecture of System

Means, manufacturers can acquire instant feedback and opportunities to get better quality of their products. Thus, mining opinions from online reviews has attracted a great deal of notice from researchers and has become an increasingly urgent activity. To analyze opinions from online reviews, it is unacceptable to simply obtain the overall sentiment about a product. In many cases, customers expect to discover fine-grained sentiments about the feature of a product that is reviewed. Opinion target is termed as the object about that which users communicate their opinions, which may be nouns or may be noun phrases also.

2. METHODS AND TERMINOLOGIES

In order to mine a data using evolutionary algorithms, it initial has to be freed from incomplete, noisy or inconsistent data. It is very important that this should be done before the start of mining process, as it will facilitate the algorithms in producing more perfect results.

If the data is taken from more than one database, they could be combined into a single database. When working with huge datasets, better result can be obtained by reducing the amount of data being handled and this could be achieved by getting a normalized sample of data from the database, resulting in much quicker result.

At this stage, the data is divided into two equal mutually exclusive elements; one is a test and the other one is a training dataset. The training dataset is used to let rules evolve which match it nearly. The test dataset will then confirm or reject these rules.

The theory of Evolutionary Algorithms (EAs) comprises of stochastic search algorithms stimulated by the process of neo-Darwinian evolution. EAs work with a population of individuals, each of them a solution to a given problem. It must be noted that this is a very generic search theory. EAs can be used to solve a wide variety of problems, through clearly specifying the type of candidate solution that an individual represents and how the quality of that solution is assessed.

Word Alignment Model (WAM) method is based on the monolingual model, which specifically mine the opinion relations between the words. Consider the phrase, “This phone has an amazing and colourful screen” as an example. Based on WAM, the opinion target and opinion word was extracted. In the above phrase, the “screen” is an opinion word and “amazing” and “colourful” is the opinion target.

When compared to earlier method syntactic patterns, the WAM precisely mine target and word. The previous nearest-neighbour method precisely mines the relation for small span sentences. But WAM method precisely mine relation for both small span and extended span relations.

The WAM method has some following constrains:

- Nouns/noun phrases should be associated with adjectives/verbs/a null word.
- Other distinct words, such as prepositions conjunctions and adverbs should be associated only with themselves.

Then the hill-climbing algorithm is used to execute local optimizations. For manipulating the relations among the words are estimated by

$$P(w_t \backslash w_o) = \frac{\text{Count}(w_t, w_o)}{\text{Count}(w_o)}$$

where, w_t represents the opinion target and w_o represents the opinion word, and then $P(w_t \backslash w_o)$ represents the problem between these two words.

Graph co-ranking is another method which is estimated with the help of candidate confidence of each opinion word and opinion target and this can be constructed with a graph. The word that has higher difficulty will be extracted as opinion word or opinion target.

3. MODIFIED WORD ALIGNMENT JOINT ASPECT AND SENTIMENT MODEL

In this section, we present our method for capturing opinion relations using unsupervised word alignment model. Similar to every sentence in reviews is replicated to generate a parallel sentence pair, and the word alignment algorithm is applied to the monolingual scenario to align a noun/noun phase with its modifiers. We select IBM-3 model as the alignment model.

$$P_{ibm}(A|S) \propto \prod_{i=1}^{n} n(\emptyset_i, w_i) \prod_{j=1}^{n} t(w_j, w_{a_j})\, d(j|a_j, n)$$

where $t(w_j|w_{aj})$ models the co-occurrence information of two words in dataset. d(j|aj , n) models word position information, which describes the probability of a word in position aj aligned with a word in position j. And $n(\varphi_i|w_i)$ describes the ability of a word for modifying (being modified by) several words. φ_i denotes the number of words that are aligned with w_i . In our experiments, we set $\varphi_i = 2$.

Since we only have interests on capturing opinion relations between words, we only pay attentions on the alignments between opinion target candidates (nouns/noun phrases) and potential opinion words (adjectives/verbs). If we directly use the alignment model, a noun (noun phrase) may align with other unrelated words, like prepositions or conjunctions and so on. Thus, we set constrains on the model:

1) Alignment links must be assigned among nouns/noun phrases, adjectives/verbs and null words. Aligning to null words means that this word has no modifier or modifies nothing.

2) Other unrelated words can only align with themselves.

Form the alignment result, alignment probability between a potential opinion target and potential opinion word is calculated using following equation.

$$P(w_t,w_o) = C(w_t,w_o)/C(w_o)$$

Similarly, we can find $P(w_t,w_o)$ by changing alignment direction in the alignment process. And then opinion association $OA((w_t,w_o)$ between $w_t$and w_o is calculated as follows.

$$OA(w_t,w_o) = (\alpha/ P(w_t,w_o)+1- \alpha/ / P(w_t,w_o))^{-1}$$

At the end confidence of the candidate is calculated using random walk with restart algorithm. Thus, we have

$$C^{k+1}=(1-\mu)\ m_{to}\ X\ C^{k}+\mu\ X\ I$$

$$C^{k+1}_{t}=(1-\mu)\ m^{T}_{to}\ X\ C^{k}_{o}+\ \mu\ X\ I^{t}_{o}$$

where, $C^{k+1}$and $C^{k+1}$are confidence of opinion target and opinion word candidate, respectively k+1 iteration, C^{k}_{t} and $C^{k}_{o}$are confidence of opinion target and opinion word candidate, respectively in k iteratarion, records opinion association among candidates, m_{ij}€ m_{to} means opinion association between ith opinion target candidates and j^{th} opinion word candidate, I and I_t are prior knowledge of candidate. Candidate with higher confidence are extracted as opinion target or opinion word.

Assumptions

We make the following assumptions about our proposed MWAJAN model:

• The generation for aspect-specific sentiments depends on the aspects. This means that we first generate latent aspects, on which we subsequently generate corresponding sentiment orientations.

• The generation for aspect terms depends on the aspects, while the generation for opinion words relies on the sentiment orientations and semantic aspects. The formulation is intuitive, for example, to generate an opinion word “beautiful”, we need to know its sentiment orientation positive and related semantic aspect appearance.

• The generation for overall ratings of reviews depends on the semantic aspect-level sentiments in the reviews. Based on the model assumptions, to generate a review document and its overall rating, we first draw hidden semantic aspects conditioned on document-specific aspect distribution. We then draw the sentiment orientations on the aspects conditioned on the per document aspect-specific sentiment distribution.

Next, we draw each opinion pair, which contains an aspect term and corresponding opinion word, conditioned on aspect and sentiment specific word distributions. We lastly draw the overall rating response based on the generated aspect and sentiment assignments in the review document. The graphical representation of the proposed method is shown in the Figure 2.

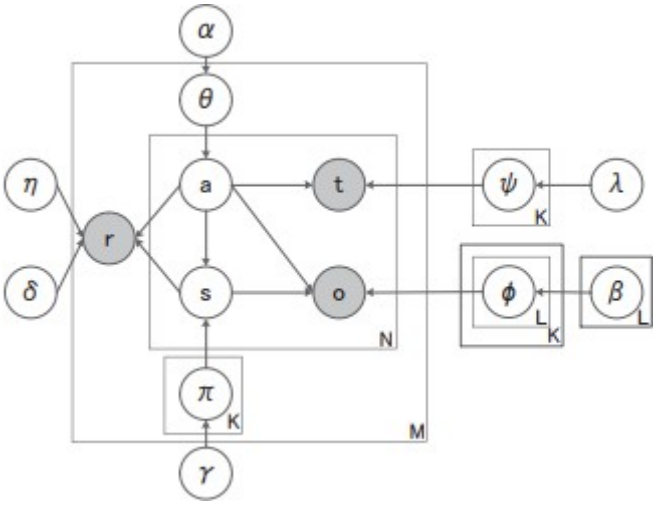

Figure 2. Graphical Representation of MWAJAM

Algorithm

The notations used in the document dm and its overall rating rm are generated from the following process:

• For each aspect k ∈ {1, . . . , K} 1) Draw aspect word distribution ψk ~ Dir(λ). 2) For each sentiment orientation l ∈ {1. . . L} a) Draw opinion word distribution ϕkl ~ Dir (βl).

• For each review d_m and its overall rating r_m

1) Draw aspect distribution Θ_m ~ Dir(α)
2) 2) For each aspect k under review r_m
a) Draw sentiment distribution n_{mk} ~ Dir(γ).
3) For an opinion pair(t_{mn},o_{mn}) n £ {1, . . . , N}
4) a) Draw aspect assignment a_{mn} ~ Mult(Θ_m).
b) Draw sentiment assignment s_{mn} ~ Mult(n_{mn}mn)
c) c) Draw aspect term mn ~ Mult(f_{mn}mn).
d) d) Draw opinion word smn ~ Mult($f_{mn}s_{mn}$)
5) Draw overall rating response rm ~ N(η T $\bar{z}$m, δ). Note that $\bar{z}$m refers to the empirical frequencies of hidden variables (latent aspects and sentiments) in the review document dm, and is defined as,

$$\bar{z} = \frac{1}{c} \sum_{n=1}^{N} (a_{mn} \text{ X } (w^t) \text{ X } (s_{mn})),$$

where ω consists of normalization coefficients on latent sentiment variables, and C means normalization constant. Under the framework of MWAJAN, overall rating response rm of review dm is drawn from a normal linear model N(η T $\bar{z}$m, δ), where η and δ refer to rating response parameters. In this normal linear model, the covariates correspond to the empirical frequencies of hidden aspects and sentiments $\bar{z}$m, and η represents the regression coefficients on the empirical frequencies.

4. RESULT

For the execution purpose of MWAJAM algorithm, java netbeans 7.4 is used as front end and mysql is used as back end. The proposed method is compared with EA and WAM and the results are tabulated in Table 1, comparing the accuracy and time period.

The Table 1 explains that EA has lower accuracy when compared to WAM and MWAJAM a takes a longer time period. Though the accuracy and time period of WAM is better when compared to EA, the proposed algorithm has a higher accuracy. The time consumed is also less in MWAJAM when compared to EA and WAM.

Table 1. Calculation of Accuracy and Time period

SL.NO.	ALGORITHM	ACCURACY	TIME PERIOD
1	EA	86.3	0.356
2	WAM	92.3	0.275
3	MWAJAM	96.78	0.214

In order to determine the efficiency of algorithm we consider another two parameters namely Precision and Recall and is compared with SP and WAM by choosing car, camera, watch, laptop and phone as test models.

Table 2 shows the precision of various algorithms by considering various test models. Table shows the recall value of various algorithms.

Table 2. Calculation of Precision by considering various products

S.No	Algorithm	Example 1 Car	Example 2 Camera	Example 3 Watch	Example 4 laptop	Example 5 Phone
1	SP	0.85	0.86	0.85	0.84	0.89
2	WAM	0.86	0.89	0.88	0.89	0.92
3	MWAJAM	0.92	0.91	0.92	0.94	0.95

Table 3. Calculation of Recall by considering various products

S.No	Algorithm	Example 1 Car	Example 2 Camera	Example 3 Watch	Example 4 laptop	Example 5 Phone
1	SP	0.86	0.87	0.75	0.83	0.89
2	WAM	0.87	0.88	0.83	0.85	0.91
3	MWAJAM	0.89	0.894	0.85	0.87	0.92

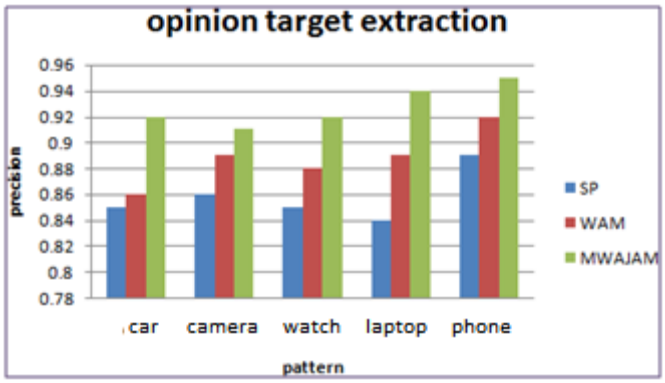

Figure 2. Opinoin Target Extraction

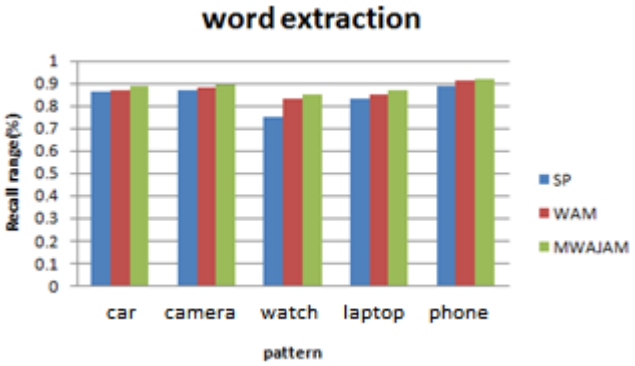

Figure 3. Word Extraction

The Figure 2 and Figure 3 clearly explains that the Precision and Recall of the proposed system is high when compared to SP and WAM. The below graph explains the opinion target extraction and opinion word extraction of SA, WAM and MWAJAM algorithm.

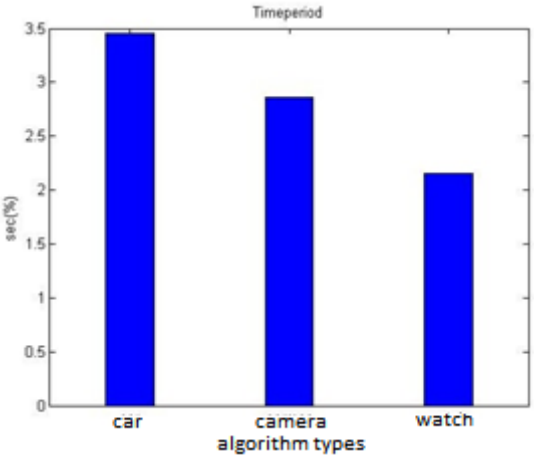

Figure 4. Time period Analysis

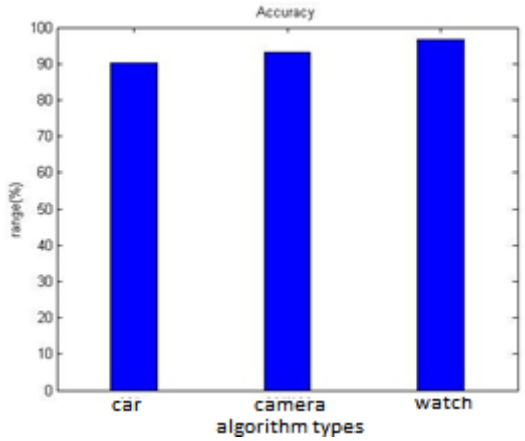

Figure 5. Accuracy Analysis

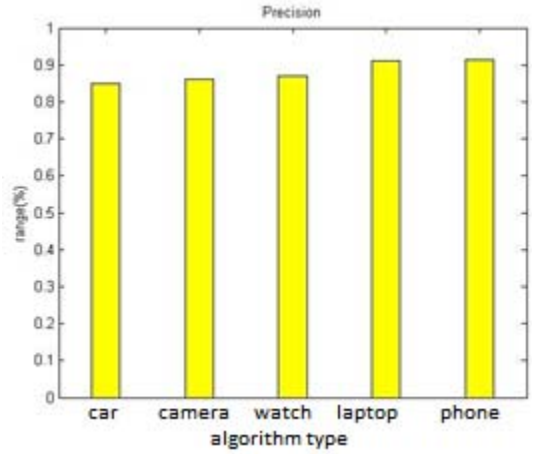

Figure 6.Precision Analysis

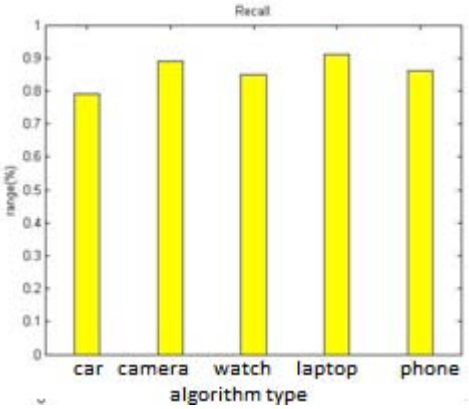

Figure 7. Recall Analysis

The parameters are analysed with various examples and is noted that the parameters vary according to various products. The Figure 4, 5, 6 and 7 explains the analysis of time period, accuracy, Precision and recall with car, camera and watch as examples.

5. CONCLUSION

The reviews about the product are necessary in-order to determine the quality of a product. The quantities of reviewers are also increasing day by day. Hence all the reviews should be reviewed quickly and should be downloaded in a short span of time without spoiling its accuracy. Thus the major challenge is extracting and comparing the data with a shorter time period and with a high accuracy. The proposed method overcomes these two major challenges when compared to previous methods.

REFERENCES

[1] K. Liu, L. Xu, and J. Zhao, "Co-Extracting Opinion Targets and Opinion Words from Online Reviews Based on the Word Alignment Model," IEEE Transactions on knowledge and data engineering, 2015, vol. 27, no. 3.

[2] Li, W. (2004). Using Genetic Algorithm for network intrusion detection. United States Department of Energy Cyber Security Group 2004 Training Conference, (pp. 24-27). Kansas City, Kansas.

[3] F. Li, S. J. Pan, O. Jin, Q. Yang, and X. Zhu, "Cross-domain co extraction of sentiment and topic lexicons," in Proc. 50th Annu. Meeting Assoc. Comput. Linguistics, Jeju, Korea, 2012, pp. 410–419.

[4] Tan, K. C., Yu, Q., & Lee, T. H. (2005). A distributed evolutionary clasifier for knowledge and discovery in data mining. IEEE Trans. on Systems, Man, and Cybernetics: Part C - Applica- tions and Reviews , 35 (2), 131-142.

[5] K. Liu, L. Xu, and J. Zhao, "Opinion target extraction using word based translation model," in Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn., Jeju, Korea, Jul. 2012, pp. 1346–1356.

[6] M. Hu and B. Liu, "Mining opinion features in customer reviews," in Proc. 19th Nat. Conf. Artif. Intell., San Jose, CA, USA, 2004, pp.755–760.

[7] G. Qiu, L. Bing, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation," Comput. Linguistics, vol. 37, no. 1, pp. 9–27, 2011.

[8] B. Wang and H. Wang, "Bootstrapping both product features and opinion words from chinese customer reviews with crossinducing," in Proc. 3rd Int. Joint Conf. Natural Lang. Process., Hyderabad, India, 2008, pp. 289–295.

[9] Terano, T., & Ishino, Y. (1996). Knowledge acquisition from questionnaire data using simulated breeding and inductive learning methods. Expert Systems with Applications , 11 (4), 507-518.

[10] T. Ma and X. Wan, "Opinion target extraction in Chinese news comments." in Proc. 23th Int. Conf. Comput. Linguistics, Beijing, China, 2010, pp. 782–790.

[11] W. X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid," in Proc. Conf. Empirical Methods Natural Lang. Process., Cambridge, MA, USA, 2010, pp. 56–65.

[12] Z. Liu, X. Chen, and M. Sun, "A simple word trigger method for social tag suggestion," in Proc. Conf. Empirical Methods Natural Lang. Process., Edinburgh, U.K., 2011, pp. 1577–1588.

[13] Q. Gao, N. Bach, and S. Vogel, "A semi-supervised word alignment algorithm with partial manual alignments," in Proc. Joint Fifth Workshop Statist. Mach. Translation MetricsMATR, Uppsala, Sweden, Jul. 2010, pp. 1–10.

[14] K. Liu, H. L. Xu, Y. Liu, and J. Zhao, "Opinion target extraction using partially-supervised word alignment model," in Proc. 23rd Int. Joint Conf. Artif. Intell., Beijing, China, 2013, pp. 2134–2140.

[15] Z. Hai, K. Chang, J.-J. Kim, and C. C. Yang, "Identifying features in opinion mining via intrinsic and extrinsic domain relevance," IEEE Trans. Knowledge Data Eng., vol. 26, no. 3, p. 623–634, 2014.

[16] J. Zhu, H. Wang, B. K. Tsou, and M. Zhu, "Multi-aspect opinion polling from textual reviews," in Proc. 18th ACM Conf. Inf Knowl. Manage., Hong Kong, 2009, pp. 1799–1802.