Sentiment Analysis and Similarity Evaluation for Heterogeneous-Domain Product Reviews

Mangal Singh
Jamia Hamdard
New Delhi,
India

Tabrez Nafis
Assistant Professor, Jamia Hamdard
New Delhi
India

Neel Mani
ADAPT Centre for Digital Content Technology
Dublin City University,
Dublin, Ireland

ABSTRACT

Sentiment analysis and classification is a prominent research topic in academics as well as in industrial field. Since each customer reviews text always state emotion about a target domain, sentiment classification is a highly domain dependent task and present study considered the reviews from heterogeneous domains. Generally researchers classify the customer review with positive, negative and neutral sentiments but a positive review can be highly positive and a negative review can be highly negative, so sentiment analysis about a review can be more effective if a sentiment scale is also defined for such greater degree of positivity or negativity. We defined a framework to classify heterogeneous product reviews with degree of polarity on a sentiment scale of range -2 to 2. For each review, an intermediate form is calculated using sentiment vectors which is further processed to calculate the sentiment polarity magnitude and similarity of reviews.

General Terms

Sentiment Analysis, Sentiment Classification, Heterogeneous Domain, Word Polarity, Review Similarity.

Keywords

Sentiment Vector, Intermediate Form, Sentiment Polarity Magnitude.

1. INTRODUCTION

Web 2.0 services enables customers to share their sentiments about the products. Due to huge size of data, automatic analysis and detection of sentiments or emotions in texts is becoming increasingly important. Product review sentiment analysis can help companies improve their products and services, and help customers make more informed decisions [7]. Analyzing customer's sentiments, opinions, evaluations and attitudes from written language is a challenging and complex task. Products are categorized in different domains and in different domains some words are used to express different sentiments for one domain and same words may convey different sentiments in other domain [4]. For example, in domain of software product reviews, the word “easy” is a positive word but for a different domain like movies or music it is considered as a negative word. Thus, a general sentiment classifier trained by combining all the labeled data from various domains may fail to capture the characteristics of each domain and cannot perform well in classifying the sentiments in a specific domain [16]. The sentiment classification is based on the recognition of sentiment carrying words in a sentence. The polarity of sentiment is identified using sentiment carrier [18]. For a product reviews, customers are also eager to find the similar reviews having similar sentiments. Different customer write their reviews with different language style but they may have same sentiments, e.g., two words "nice" and "good" represents same sentiments.

By considering above observation, in this approach authors are splitting sentiment training data in two layers, i.e., a generic training data and a domain specific training data. Also each entity in training data is containing words (sentiment carriers) and a specified sentiment vector. Multiple words represents same sentiments will have same sentiment vector. Each sentiment vector is defined with polarity and strength. A sentiment scale is defined by minimum and maximum strength from the set of sentiment vectors. The training data is having limited set of words and is referenced to build vocabulary using WordNet [20, 21], a large lexical database that also provides relationship information among words and concepts. Semantic similarity between words such as nouns, verbs and adjectives can be easily evaluated using WordNet. Customer use short forms, emotion symbols and other irregularities in writing reviews. Text pre-processing techniques are used to standardize certain tokens of review’s text and to increase the accuracy of analysis [5]. Pre-processed text of the review is further processed and an intermediate form of review is calculated by replacing sentiment carriers with respective sentiment vectors. Two reviews "Product is Good" and "Product is Nice" are similar in terms of sentiment expression and since same sentiment vectors are being used for "Good" and "Nice", both reviews will have same intermediate form. For each review, intermediate form is used to calculate the sentiment polarity magnitude and to find co-ordinate position of review on the sentiment scale.

2. RELATED WORK

This section contains the review of representative works related to sentiment analysis, heterogeneous-domain sentiment classification and document similarity. One major challenge in sentiment analysis is to handle irregularities in language of text. Customer reviews are generally having the varying and unpredictable nature of language; it is likely that preprocessing techniques could be used to standardize certain tokens of reviews text [5]. Some researchers have put stress on text pre-processing and they used different text pre-processing techniques [5, 6]. Here we used some techniques as Replace Emotion Symbol, Upper Case Identification, Word Compression, Word Segmentation [6] and Stop Word Removal [5]. Further each customer review text is represented as a continuous attributes and its analysis is complex due to such larger degree of attribute dimensions. Chun-Han Chu *et al.* has focused on word polarity classification, which is extended to perform classification of sentences and paragraphs [18]. In their work, a semantic class labeler is based on sentiment sensitive vector for different POSs and polarities. In different domains different words are used to express sentiments, and the same word may convey different sentiments in different domains [4]. To address the problem of multi-domain sentiment classification [1] has used two types of classifiers, a general sentiment classifier and a domain specific sentiment classifier. Bollegala *et al.* has modeled sentiment classification as the problem of training a binary classifier using reviews annotated for positive or negative sentiment and also create a sentiment sensitive distributional thesaurus using labeled data for the source domain and unlabeled data for both source and target domains [3]. They also incorporated document level sentiment labels in the context vectors as the basis for measuring the distributional similarity between words. A method for calculating semantic similarities between document is given [12] and explained that the overall similarity between documents is a combination of cosine similarity and semantic similarity. To calculate semantic similarities between documents, they proposed a method which is based on cosine similarity calculation between concept vectors of documents obtained from taxonomy of works that capture IS-A relations [12].

3. METHODOLOGY

In this approach, a framework has been purposed for sentiment analysis of customer reviews by considering various factors that affect the sentiment analysis the most. New method used by constructing an intermediate form for each review by using sentiment vectors. The calculation of sentiment polarity magnitude and proximity analysis among reviews is based on the intermediate form.

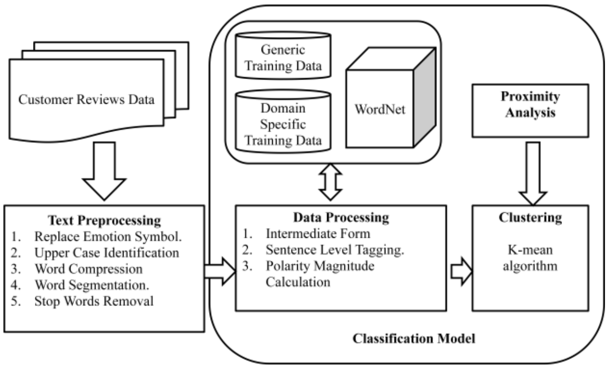

Fig 1: Proposed Framework.

3.1 Text Preprocessing

Customer uses various irregularities in language while writing their reviews. Text of reviews may have emotion symbols, upper case words to stress their emotion, last letter repeating to greater degree of emotion, intensifier and negative words before adjectives and some words that do not express any sentiments. Customer reviews are preprocessed to clean text for further processing. The techniques adopted in proposed framework are used commonly in information retrieval applications.

3.2 Sentiment Vectors

Each word (Sentiment Carrier) is either positive, negative or neutral, and thus a sentiment vector is assigned to each word in training data. Each sentiment vector is having two attributes as polarity and strength (→v[P, S]).

Polarity ← [P: Positive, N: Negative, U: Neutral]
Strength ← [-2,-1,0,+1,+2] .

3.2.1 Sentiment Scale

Sentiment scale also depends upon the set of sentiment vectors and the range of sentiment scale is between the maximum and minimum strength of the sentiment vectors. Here the maximum strength is 2 and minimum strength is -2, thus the sentiment scale is ranging between 2 and -2. Thus, sentiment scale is automatically designed with the design of the set of sentiment vectors.

3.3 Domain Specific Training Data

Two layered training data are being used as generic training data layer and domain specific training data layer. Generic training data layer contains general words having same sentiments for all domains. Domain specific layer contains words which express different sentiments for different domains. Set of sentiment vectors is designed with human intelligence and each entry in training data is assigned with corresponding sentiment vector so there is a 1:M relationship among sentiment vectors and words.

3.4 Building Emotion Vocabulary

This step involves creating a vocabulary with words annotated with sentiment vectors. The vocabulary is then searched to find the respective sentiment vectors of words to construct the intermediate form of review text for further processing. Semantic similarity between words like nouns and verbs can easily evaluated using WordNet and thus it is used to build emotion vocabulary. Each entry in emotion vocabulary is also mapped with a corresponding sentiment vector and during processing these words (sentiment carriers) are replaced with the sentiment vector to generate native intermediate form for each review text.

3.5 Data Processing

Data processing started with lexical analysis and tokenize each review. Relative sentiment vectors are assigned to the various terms, and therefore support a semantic-driven proximity in the feature space of each review. Each processed review is converted to a native intermediate form, which contains only the sentiment vectors.

3.5.1 Sentence Level Tagging

User reviews may have multiple sentences and each sentence may express different kind of emotion. Here each review is considered as an array of sentences. With lexical analysis and by referencing vocabulary, each review is transformed to intermediate form. If we are having *n* number of reviews and *r* is a review from set of reviews *R* then, *intermediate_r* is the intermediate form of review *r*. Sentence level tagging is a step to calculate the polarity and its magnitude for each sentence. If a sentence s of intermediate form *intermediate_r* is having *m* words then the Sentence Level Polarity Magnitude (*m_SLT*) and sentence level polarity (*p_SLT*) will be calculated as follows:

$$m\_SLT = \frac{\sum_{i=1}^{m} \rightarrow v(i)}{m} \qquad (1)$$

$$p\_SLT = \begin{cases} P, if\ m\_SLT\ >\ 0 \\ N, if\ m\_SLT\ <\ 0 \\ U, if\ m\_SLT\ =\ 0 \end{cases} \qquad (2)$$

3.5.2 Sentiment Polarity Magnitude

For Each review, user sentiment can be positive, negative or neutral but here we are more interested in calculating the magnitude or degree of positivity and negativity of each review. Final polarity magnitude of a review is calculated by adding the polarity score of each sentence and dividing it to number of sentences. Since review *r* is having *n* sentence, the Sentiment Polarity Magnitude (*M*) is the average of *m_SLT*.

$$M = \frac{\sum_{i=1}^{n} m\_SLT\,(i)}{n} \qquad (3)$$

3.5.3 Clustering & Proximity Analysis

User specified numbers of cluster are built based on the sentiment polarity magnitude of reviews. Proximity analysis is performed among the reviews within the same cluster to find the similar reviews as well as to study the average similarity among reviews with varying number of clusters. K-means clustering algorithm is used to define clusters.

The angle between two vectors (reviews) is measured by cosine similarity the and it is calculated with an assumption that words having same sentiment vector shows same sentiments and thus similar to each other. Therefore instead of using the actual review text, here we consider intermediate form of reviews. If *r1*, *r2* are two reviews from review set *R* and *intermediate_r1*, *intermediate_r2* are respective intermediate forms then based on above theory, cosine similarity (*simCOS*) of *r1* and *r2* will be equal to cosine similarity of *intermediate_r1* and *intermediate_r2* in terms of sentiment expression.

simCOS(*r1*, *r2*)

$$= \frac{r1\ .\ r2}{||r1\,||\,.||r2\,||}$$

$$= simCOS\ (intermediate\_r1, intermediate\_r2)$$

$$= \frac{intermediate\_r1\ .intermediate\_r2}{||\ intermediate\_r1\ ||\ .||\ intermediate\_r2\ ||} \qquad (4)$$

If there are *n* reviews in a cluster then, average cosine similarity of review *r1* with other reviews in same cluster is:

$$avg\_simCOS(r1) = \frac{\sum_{j=2}^{n-1} sim\,COS\,(r1,rj\,)}{n} \qquad (5)$$

Average cosine similarity among *n* reviews within a cluster *C* is:

$$avg\_simCOS(C) = \frac{\sum_{j=1}^{n} avg\_sim\,COS\,(ri)\,)}{n} \qquad (6)$$

4. EXPERIMENT

Amazon™ reviews database is downloaded for experiment and analyzed 1000 positive and negative reviews for each domain (Software, DVD & Music). This database is in XML format and is imported to MS SQL Server 2012. Two layered training data set is used. One Generic layer contains the common words for all domains and here for this experiment we used 520 words for generic layer and 100 words each for domain specific layer. A software application is developed based on the theory explained earlier. Software is written in C#.Net (UI) with MS SQL Server 2012(Database). This software is capable of importing candidate data from XML format to MS SQL Server and processing the reviews text. Since the review's title also expresses the reviewer sentiment, it is also merged with review text for processing. A good degree of accuracy is observed in results as shown in Table 1.

Table 1: Sentiment Analysis

Review Type	%age of Accuracy
Software Positive	78.9%
Software Negative	80.5%
Music Positive	82.6%
Music Negative	77.9%
DVD Positive	83.8%
DVD Negative	78.4%

All reviews are arranged in shorted order based on the sentiment polarity magnitude and then placed on the sentiment scale of -2 to 2. In Fig 2, top three layers are for positive and bottom three are for negative reviews.

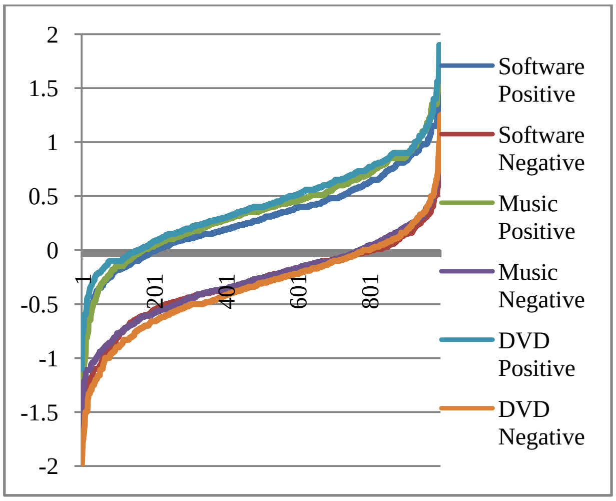

Fig. 2 Sentiment Scaling

Average cosine similarity is calculated within a cluster by considering intermediate forms of reviews. Similarity study is done with varying number of clusters and following results are found. With increase in numbers of clusters the average cosine similarity also increased. Since sentiment data mining is about to find the positivity or negativity of reviews, to calculate cosine similarity the neutral sentiment vectors are not considered.

Table 2: Average Cosine Similarity

No. of Clusters	Average Cosine Similarity
2	0.49
5	0.77
10	0.93

5. CONCLUSIONS

In present work, we demonstrated sentiment classification and scaling with similarity evaluation among reviews. Review data is pre-processed and cleaned for data processing. Multi layered training data and related sentiment vectors with WordNet are used to transform reviews to intermediate form. Since the only interest is in sentiments not in the language, the crux of present theory is based on the intermediate form. Sentiment polarity score and semantic cosine similarity is computed based on the intermediate form of each review and similar reviews are identified which are more identical in terms of sentiments or emotion. A comparative study was also made among the varying numbers of clusters and similarity among the reviews in each cluster.

6. FUTURE WORK

For sentiment classification, we tried with a composite approach by considering many factors but during processing it is observed that sometimes customer writes reviews about a product in comparison to other products. So here the biggest challenge is to find the subject of the speech. This study can be further extended using natural language processing to handle such comparison.

7. REFERENCES

[1] Wu, F. and Huang, Y. 2015. "Collaborative Multi-domain Sentiment Classification," Data Mining (ICDM), IEEE International Conference on, Atlantic City, NJ, pp. 459-468.

[2] Bisio, F., Gastaldo, P., Peretti, C., Zunino, R. and Cambria, E. 2013. "Data intensive review mining for sentiment classification across heterogeneous domains," Advances in Social Networks Analysis and Mining (ASONAM), IEEE/ACM International Conference on, Niagara Falls, ON, pp. 1061-1067.

[3] Bollegala, D., Weir D. and Carroll, J. 2013. "Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus," in IEEE Transactions on Knowledge and Data Engineering, 25(8):1719-1731.

[4] Glorot, X., Bordes, A. and Bengio, Y. 2011. “Domain adaptation for large scale sentiment classification: A deep learning approach”. In proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 513-520

[5] Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N. and Perera, A. 2012. "Opinion mining and sentiment analysis on a Twitter data stream," Advances in ICT for Emerging Regions (ICTer), International Conference on, Colombo, pp. 182-188

[6] Nie, P., Zhao, X., Yu, L., Wang, C. and Zhang, Y. 2015. “Social Emotion Analysis System for Online News”. In proceedings of 12th Web Information System and Application Conference.

[7] Hu, M. and Liu, B. 2004. Mining and summarizing customer reviews. In KDD, ACM, pp 168-177.

[8] Pang, B. and Lee, L. 2008. Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2):1-135.

[9] Ren, F. and Wu, Y. 2013. Predicting user-topic opinions in twitter with social and topical context. IEEE Transactions on Affective Computing, 4(4):412–424.

[10] Lin, D. 1998. An information-theoretic definition of similarity. In Proceedings of 15th International Conference on Machine Learning. Morgan Kaufmann, San Francisco, CA, pp. 296–304.

[11] Zhou, Y. 2015. “The analysis of online users’ emotions based on data mining”, 3rd International Conference on Machinery, Materials and Information Technology Applications (ICMMITA 2015).

[12] Madylova, A. and Oguducu, S. G. 2009. “A Taxonomy based Semantic Similarity of Documents using the Cosine Measure,” IEEE.

[13] Tang, D., Wei, F,. Qin, B., Yang, N., Liu, T. and Zhou, M. 2016. "Sentiment Embeddings with Applications to Sentiment Analysis," in IEEE Transactions on Knowledge and Data Engineering, 28(2): 496-509.

[14] Glorot, X., Bordes, A. and Bengio, Y. 2011. “Domain adaptation for large scale sentiment classification: A deep learning approach,” ICML.

[15] Lin, L., Jianxin, L., Zhang, W. and Sun, Y. 2014. Opinion Mining and Sentiment Analysis in Social Networks: A Retweeting Structure-Aware Approach. In proceedings of IEEE/ACM 7th international Conference on Utility and Cloud Computing, Washington, DC, USA,

[16] Blitzer, J., Dredze, M. and Pereira, F. 2007. “Biographies, Bollywoo, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification.” ACL, 7: 440-447.

[17] Pang, B., Lee, L. and Vaithyanathan, S. 2002. "Thumbs up?: sentiment classification using machine learning techniques." ACL, pp.79-86.

[18] Chun-Han Chu, Apoorva Honnegowda Roopa, Yung-Chun Chan, and Wen-Lian Hsu. 2015. "Constructing sentiment sensitive vectors for word polarity classification." In proceedings of Conference on Technologies and Applications of Artificial Intelligence, pp. 252-259.

[19] Balla-Muller Nora, Lemnaru, C. and Potoles, R. 2010. “Semi-Supervised Learning with Lexical Knowledge for Opinion Mining”. In Proceedings of IEEE 6th International Conference on Intelligent Computer Communication and Processing, pp. 19-25.

[20] George A. Miller. 1995. WordNet: A Lexical Database for English. Communications of the ACM, 38(11): 39-41.

[21] Fellbaum, C. 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press. 422 p.

[22] McAuley, J., Targett, C., Shi, Q. and Hengel, A. van den. 2015. “Image-Based Recommendations on Styles and Substitutes”. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 43-52.

[23] McAuley, J., Pandey, R. and Leskovec, J. 2015. “Inferring networks of substitutable and complementary products”. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794.