A Dependency-Directed Opinion Analytics For Product Review Classification Based On Keyphrase

Dhanasekaran K, Manikandan Ramasamy, Raju Shanmugam, M Prathilothamai

Abstract: Text classification on product reviews has long been a challenging task due to the rapid growth of Web usage that has resulted in a huge volume of unstructured data. Recently, Opinion mining has been emerged as an important discipline to process the unstructured data. Although several opinion mining approaches addressed the problem of dealing with unstructured data, further research opportunities are available due to the issues like class imbalance, and complexity in text data analytics that affects the performance of opinion learning. Further, the manual text classification consumes a lot of time while identifying useful information. Also, the existing approaches for classifying texts based on majority category are not enough for realistic scenarios specifically in large scale applications. This paper proposes a prediction approach which focuses on obtaining useful information by using keyphrase and category labels. In this paper, we first investigate existing machine learning techniques to classify customer opinions with respect to multiple categories. Moreover, we propose keyphrase based multiclass text classification that finds insights from opinions of various customers on financial products and services. The result of our experiment shows that our dependency-directed opinion learning can show significant improvement over precision, recall, and F1-measure.

Index Terms: Product review classification, opinion mining, multiclass text classification, machine learning.

1 INTRODUCTION

The rapid development in industry and digital technologies has increased availability of information. However, only a few attempts have been made on opinion learning related to customer opinion analytics. The main motivation behind the use of opinion learning techniques in customer review analytics is that it can discover hidden patterns from training data. It also provides robust results on test patterns. In behaviour prediction, analyzing the behaviour of various objects to detect object interaction play a vital role in big data organization [1] and it helps in developing a large database or repository. Moreover, technological development, and Internet availability increase the amount of text data and text documents. Therefore, the manual information extraction becomes a difficult task [2]. Further, the actual processes change dynamically. Hence, the method for developing a large database is needed to acquire current knowledge. Nowadays, the availability of mobile apps has increased the access to social network sites which increases the amount of user comments day by day. Automated classification of user comments has become an important application for improving the quality of product and services in financial domain. The text classification involves data analysis components like co-occurrence based data analytics and rule-based analytics. In co-occurrence based approach, co-occurrence among names of objects is considered as important information for representing useful relationships.

- *Dhanasekaran K is currently working as Associate Professor in Department of Computer Science and Engineering, Jain College of Engineering, VTU, Belagavi, India. E-mail: jcedhana16@gmail.com*
- *Manikandan Ramasamy is currently working as a Senior Assistant Professor in School of Computing Science and Engineering, VIT Bhopal University, India. E-mail: clickmani@gmail.com*
- *Raju Shanmugam is currently one of the professors and Dean in School of Computing Science and Engineering, Galgotias University, Uttar Pradesh, India. E-mail: srajuhere@gmail.com*
- *Prathilothamai M is is currently working as a Senior Assistant Professor in Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India, E-mail: m_prathilothamai@cb.amrita.edu*

Whereas, in rule-based approaches, predefined rules or patterns have been used, so, these approaches can access only known patterns. It cannot find interesting information from new or unknown interaction patterns. The interaction pattern in different domains varies and has its own pattern with respect to the individual domain. The interesting property of machine learning techniques is that it can discover the hidden patterns from training data. It also provides robust results on test patterns. In this paper, product review classification aimed to predict the opinions related to various product categories. The unigrams and bigrams analysis is performed to evaluate the patterns present in the textual data. The opinion grouping according to the product categories supports natural language processing at the classification level. The problem addressed in our paper is formulated as a supervised text data analytics for improving performance of customer opinion analytics. Proposed approach attempts to find useful opinion by using linear support vectors, logistic regression, random forest classification, through dependency-directed opinion learning. It shows high precision on review classification task when linear support vector machine performs opinion prediction for different categories. When a new opinion comes in, that will be assigned to one of 12 categories that we considered for data analytics.

2 LITERATURE REVIEW

Due to the rapid development of Web, a huge amount of text data is being generated every day. It also allows users to view information with respect to some classes which will be helpful for customers' review classification [3]. A practical approach discussed in [4] has used twitter data to gain insights to promote physical activity. This approach has several limitations. One of the limitations is that it cannot generalize well as a result of user characteristics mainly because of adult behaviors and the use of diverse linguistic structures. Other issues include unknown values and each tweet has different probability of occurrence. Applying sentiment analysis to tweets has become an important research area because of the huge amounts of user generated tweets. This research effort

has explored a hypothesis by utilizing locations, times, and authors [5]. Further, a Bayesian approach has been used to combine variables. In sentiment analysis, public opinion is used to capture important information. In this perspective, linking of public opinion to public sentiment has been focused in [6]; it has utilized text streams to find correlation between sentiment word frequencies and opinions. Nowadays, social networks provide a way to share opinion about products. The sentiment analysis discussed in [7] has used distant supervision and machine learning algorithms to classify Twitter messages. This survey work has used Parts-of-Speech (POS)-specific prior polarity features and a tree kernel has been used to prevent the need for monotonous feature engineering. Sentiment analysis has been emerged as an important technique for predicting the result of election. Even though many researches are going on in opinion mining, still there are research gaps to develop suitable opinion mining approach. In this perspective, a comparison of sentiment analysis techniques presented in [8] has applied Support Vector Machine (SVM) to analyze political views. Text mining used in [9] has focused on analyzing Twitter accounts data. This research study has analyzed the tweets to identify topics and variations over time. Further, it performed social network analysis to analyze followers related to immigration and citizenship based on new legislations and policies. Finding interesting insights from customer opinions related to products is an important task for improving market analysis. A methodology discussed in [10] deals with interpretation issue for analyzing Twitter data. In order to analyze user interests, the majority of existing work performs social network analysis [11]. For behaviour prediction, accounts that follow users have been examined by using some set of features that describes the user profiles. The accuracy of prediction of target attribute-value will decrease if only a limited amount of posts is available. The research method which uses likes for topic of discussion can significantly improve the performance of prediction. Most of the social network online services allow users to associate with others by own or else it allows users to suggest posts to other friends. In this case, social network is represented like graph [12], where users are nodes, and edges denote the association between nodes. Given the areas of user, a recommender system can provide a way to infer user demographic traits; it can help in predicting the groups that the user likes [13]. The technique like labelling users with respect to political leader they follow will lead to sampling biases [14] and it reduces performance of training models when it applied to users different from some kind of population. Furthermore, asking user to answer set of questions is a time consuming process [15]. As mentioned in [16], the opinion mining can be modeled as topic-sentiment model if there is enough training data. For Twitter sentiment classification, a distant supervision has been utilized in [17]. It collects labeled tweets at different times from different locations. Further, analysis of tweet sentiments is performed across some categories. According to the co-occurrences of terms in tweets, a set of linked terms has been identified using term document matrix [18]. In sentiment analysis, the training data may consist of emoticons, and acronyms. These data are expressed as noisy labels in [19]. Also, the messages containing negation needs to be considered [20] as it affects polarity and improves the accuracy of the classification. Recently, a document clustering approach proposed in [21] used graph, and it considered PubMed data. It did not apply natural language processing. This piece of research work applied interactive graph layout algorithm for data visualization. The E-Coli and yeast dataset are considered to find insights from clusters and documents. Further, Text data mining researchers focus on identification of gene patterns from text data. Authors in [22] created a dictionary for identifying chemical names. The dictionary developed with an aim to use for health risk analysis and food safety analysis. Furthermore, the major tasks like rule-based filtering, frequent pattern analysis, and disambiguation have been implemented for creating dictionary. Authors has compared combined dictionary with derived dictionary which is developed using an annotated corpus. This dictionary-based approach indicated that pre-processing of terms, using limited manual checking for frequent patterns, along with some disambiguation can improve precision, with less loss of recall. Further, the combined dictionary was shown to be better than standard dictionary for achieving better performance.During the past few years, several efforts were made by researchers on semantic role labelling which aims to extract hidden semantics of terms. In paper [23] evaluation of domain adaptation technique for semantic role labelling has been carried out in biomedical domain. The bio-computing system has implemented semantic role labelling.Authors in [24] proposed DNorm to perform disease normalization, which finds similarities between disease names and concept names. The interoperability issue that occurs on inter-process communications has been addressed in [25]. To deal with that issue, instead of integrating named entity recognition (NER), HTTP request from text mining component can be of help in providing Web services to identify biological terms.Interestingly, causal networks are useful for gene-expression analysis. Because, this network finds hypothesis which highlights changes occurring in expression. A causal network discussed in [26] performs pathway analysis. It mainly works on upstream data. Authors also focused on extending the method to predict downstream effects on disease, and biological functions. Recent study aimed to detect tumor cells while performing cancer diagnosis. THetA2 algorithm in [27] attempted to find the number, and content of tumor subpopulations directly, from DNA sequences. Another study has focused on protein function prediction. This provides approximation over gene sequences [28]. An AkaneRE system produced filtered protein interactions but it used interactions which are manually generated. AkaneRE system uses an XML configuration file, and publicly available corpora [29]. A protein target prediction captures structural features through integrated system. It tried to assess side effects, and drug repositioning [30]. A framework that uses neighborhood analysis proposed to predict functional annotation using gene interaction networks [31]. Authors in [32] applied the majority voting technique to predict cancer class through the rules of genetic programming. Moreover, it discussed that some frequent genes in the rules of GP classifier acts as the potential biomarkers for certain cancers.

3 METHODOLOGY

3.1 Dependency-Directed Opinion Classification

The proposed dependency-directed opinion learning focused on examining training examples to find insights from text data. The pre-processing reads the input text data to remove errors and other inconsistent data which further helps in optimal

concept approximation to calculate number of consistent opinions. The missing values and duplicate data are removed to increase the quality of dataset for further opinion learning task. In our case, the number of opinions per product category cause data imbalance problem, hence, conventional machine learning algorithms is biased towards the major categories. They do not take the data distribution into account and minority categories are treated as outliers. In some cases, we need to artificially balance the dataset by using sampling techniques. However, in our proposed opinion learning, the majority categories are treated as interesting target for opinion classification with reasonable prediction accuracy.

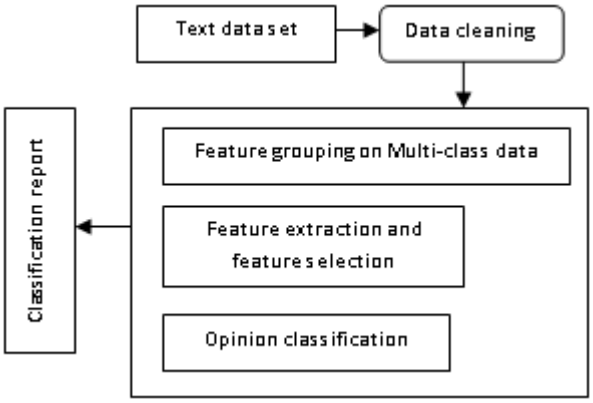

Fig.1 *Architecture of Opinion Analytics*

The text classifier cannot directly learn from the textual data. Therefore, during the preprocessing, the numerical feature vectors are generated by extracting features from text. To remove the noisy features, all stopwords were removed while computing the number of occurrence of words in a document. For each term in the opinion text, term frequency and inverse Document frequency (tf.idf weighted vectors) is calculated through TfidfVectorizer of sklearn package. We represented the tf.idf score for different unigrams and bigrams features. After the data transformation, we had all the features and labels to train the multi-label opinion classification models. We applied Logistic regression, Linear SVM, NB Classifier and Random forest classification for experimenting with different text classification models. Our proposed opinion classification approach is shown in Figure 1.

4 RESULTS AND DISCUSSION

To ensure consistency of classification, we evaluated the prediction accuracy and analyzed the reason for analytical issues. In our case, linear support vector machine and logistic regression has shown significant improvement over other classification models. Using the best linear support vector model, the confusion matrix is constructed and identified the discrepancies between predicted labels and actual labels. The majority of the predictions appeared on the diagonal. However, there are a number of misclassifications which can be further addressed by analyzing the possible causes. Some of the misclassified opinions support more than one category. The credit risk related opinion grouped towards both credit card and credit report. In order to handle the misclassification issue, the chi-squared test was used and the most correlated terms to each of the categories are generated as a result of it. The final opinion classification report for each class is shown in Table 1. For instance, the classification model predicted debit collection as credit reporting and credit repair services. Let t(x) be the target attribute of instance x. Opinion learning predicts the opinion that belongs to the target attribute-value satisfying the necessary constraints. In this work, a text document is represented by classes, terms, and object-interactions based weights. The terms with higher weights have more discriminative power than terms with lower weights. When a term appears many more times in one specific class, it gets higher weight. If it appears more times in all other classes, the weight gets reduced. The vector space represents a term document matrix (tdm) that represents a document collection. Initially, let d be a document represented in the vector space as follows:

$$f(d)= [tf(t1,d),tf(t2,d),...,(t_D,d) \epsilon R^D$$

Here, D is the size of the dataset of the corpus, and tf(ti,d) denotes the frequency of term ti in document d. The function f represents d as a term frequency vector. In order to improve the bags-of-word representation with semantic information, the matrix M is created using the object-interaction weighting approach. In this matrix, i, and j represents element of M quantitatively considering the interaction information between term ti and tj. The proposed method takes advantage of these calculated weights while mapping is done. It is represented by M=SST, where S is a semantically rich object-interaction matrix. This matrix M is used to transform input documents into the feature space. The information stored in the feature space is analyzed to predict the target class of test pattern. Proposed method shows significant improvement over precision rate for opinions of high probability. Further, this method can make use of additional data to improve the quality of text data analytics. And, the natural language parser is somewhat slow in interaction processing. Therefore, there is further opportunity for improving speed through better parsing and pre-processing approach, which is to be tried in our future work.

Table 1 *Classification result*

Performance metrics	Precision	Recall	f1-score	Support
Credit card	0.5	0.40	0.45	27
Credit reports	0.63	0.84	0.72	191
Debt collection	0.73	0.84	0.78	164
Prepaid card	0.38	0.32	0.35	41
Mortgage	0.72	0.86	0.79	87
Bank service reports	0.42	0.29	0.34	28
Credit reporting	0.56	0.16	0.25	57
Student loan	0.87	0.7	0.78	37
SB account	0.45	0.57	0.50	23
Vehicle loan	0	0	0	10
Consumer Loan	0.57	0.29	0.38	14
Online transaction services	1	0.2	0.33	10
Money transfers	0	0	0	3
Payday loan	0	0	0	1
Personal loan	0	0	0	9

Financial service	0	0	0	1
Prepaid card	0	0	0	3
avg/total	0.62	0.65	0.61	706

5 CONCLUSION

The proposed dependency-directed opinion learning predicts useful opinions incorporating the methods like data cleaning, feature extraction (tfidfvectorizer), feature selection (chisquare) and opinion classification for data analytics. This approach with the support of co-occurrence analysis classifies opinions for multiple categories. Unlike other opinion mining methods that use keyword matching and predefined information, proposed approach shows significant improvement over prediction accuracy through dependency-directed random forest, linear support vector machine, and logistic regression. And, the natural language parser is somewhat slow in interaction processing, therefore, further analysis for improving speed through efficient parsing and pre-processing are to be tried in our future work. The proposed method successfully performed text data analytics, which analyzes object names, and action terms that are used for entity and interaction analytics. The majority of the texts in the dataset were successfully processed. However, efficient pre-processing is necessary while dealing with large sized test data set which contains large amounts of textual content, that will help achieve the reasonable results on complex test data.

6 ACKNOWLEDGMENT

Author would like to thank Prof. K. Nallathambi, Professor (Rtd.,), GCE, Salem, Dr. K. G. Vishwanath, Principal & Director, JCE, Belagavi, and Prof. Udhaya Chandra, Senior Director, JCE, Belagavi, for their consistent encouragement and support.

REFERENCES

[1]. Ali W, Rito T, Reinert G, Sun F, Deane CM (2014) Alignment-free protein interaction network comparison. Bioinformatics30:i430-i437.

[2]. Jamieson DG, Gerner M, Sarafraz F, Nenadic G, Robertson DL (2012) Towards semi-automated curation: using text mining to recreate the HIV-1, human protein interaction database. Database2012: p.bas023.

[3]. G. Geetika, and Y.Divakar, Sentiment Analysis of Twitter Data using Machine learning approaches and semantic analysis, IEEE Conference paper, DOI: 10.1109/IC3.2014.6897213, 2014.

[4]. Y. Sunmoo, E.Noemie, and B.Suzanne, A practical approach for content mining of tweets, American Journal of Preventive Medicine, vol.45, pp.122-129, 2014.

[5]. V.Soroush, Z.Helen, R.Deb, Enhanced twitter sentiment classification using contextual information,In Proceedings of WASSA, pp.16-24, 2015.

[6]. C.Brendan, B.Ramnath, R.R, Bryan, A.S. Noah, From tweets to polls: linking text sentiment to public opinion time series,In Proceedings of the fourth international conference on Weblogs and Social Media, pp.122-129, 2010.

[7]. S.Varsha, S.Vijaya, P.Apashabi, Sentiment analysis on twitter data, International Journal of Innovative Research in Advanced Engineering, vol.2, pp.178-183, 2015.

[8]. H.Ali, M.Sana, K.Ahmad, and S.Shahaboddin, Machine learning-based sentiment analysis for twitter accounts, Mathematical and Computational Applications, vol.23, pp.1-15, 2018.

[9]. Z.Yanchang, Analysing twitter data with text mining and social network analysis, In Proceedings of AusDM, pp.41-47, 2013.

[10]. A.H. Syed Akib, M.TahmidEkram, I.MohammadSamiul, A.Faysal, and M.R. Rashedur, Localized twitter opinion mining using sentiment analysis, Decision analytics, vol.8, pp.1-19, 2015.

[11]. V.Svitlana, B.Yoram, V.D. Benjamin, Mining user interests to predict perceived psycho-demographic traits on twitter,In IEEE Second International conference on Big data computing service and applications, 2016.

[12]. S. Volkova, G. Coppersmith, and B. Van Durme, Inferring user political preferences from streaming communications, In Proceedings of ACL, pp. 186–196, 2014.

[13]. P. Kapanipathi, P. Jain, C. Venkataramani, and A. Sheth, User interests identification on twitter using a hierarchical knowledge base in The Semantic Web: Trends and Challenges, Springer, pp. 99–113, 2014.

[14]. R. Cohen and D. Ruths, Classifying political orientation on Twitter: It's not easy!,In Proceedings of ICWSM, 2013.

[15]. H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, R. E.Lucas, M. Agrawal, G. J. Park, S. K. Lakshmikanth, S. Jha, M. E.Seligman et al., Characterizing geographic variation in well-being using tweets,In Proceedings of ICWSM, 2013.

[16]. Q.Mei, X.Ling, M.Wondra, H.Su, and C.X. Zhai, Topic sentiment mixture: modeling facets and opinions in weblogs, In Proceedings of the 16th International conference on WWW, 2007.

[17]. G. Alec, H.Lei, and B.Richa, Twitter sentiment analysis, Entropy, 17, 2009.

[18]. Gentry et. Al., Rgraphviz: Provides plotting capabilities for R graph objects, R package version 2.4.1, 2013.

[19]. B.Luciano, and F.Junlan, Robust sentiment detection on twitter from biased and noisy data,In Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 36-44, 2010.

[20]. G.Vinodhini, R. M. Chandrasekaran, Sentiment Analysis and Opinion Mining: A Survey, IEEE, vol.2, 2012.

[21]. Theodosiou T, Darzentas N, Angelis L, Ouzounis CA (2008) PuReD-MCL: a graph-based PubMed document clustering methodology. Bioinformatics24:1935-1941.

[22]. Hettne KM, Stierum RH, Schuemie MJ, Hendriksen PJ, Schijvenaars BJ, Van Mulligen EM, Kleinjans J, Kors JA (2009) A dictionary to identify small molecules and drugs in free text. Bioinformatics25:2983-2991.

[23]. Dahlmeier D, Ng HT (2010) Domain adaptation for semantic role labeling in the biomedical domain. Bioinformatics26:1098-1104.
[24]. Leaman R, Doğan RI, Lu Z (2013) DNorm: disease name normalization with pairwise learning to rank. Bioinformatics p.btt474.
[25]. Wiegers TC, Davis AP, Mattingly CJ (2014) Web services-based text-mining demonstrates broad impacts for interoperability and process simplification. Database,2014: p.bau050.
[26]. Krämer A, Green J, Pollard J, Tugendreich S (2013) Causal analysis approaches in ingenuity pathway analysis (ipa). Bioinformatics p.btt703.
[27]. Oesper L, Satas G, Raphael BJ (2014) Quantifyingtumor heterogeneity in whole-genome and whole-exome sequencing data. Bioinformatics30:3532-3540.
[28]. Aßhauer KP, Wemheuer B, Daniel R, Meinicke P (2015) Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. Bioinformatics31:2882-2884.
[29]. Sætre R, Yoshida K, Miwa M, Matsuzaki T, Kano Y, Tsujii JI (2010) Extracting protein interactions from text with the unified AkaneRE event extraction system. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)7:442-453.
[30]. Naveed H, Hameed US, Harrus D, Bourguet W, Arold ST, Gao X (2015) An integrated structure-and system-based framework to identify new targets of metabolites and known drugs. Bioinformatics p.btv477.
[31]. Bogdanov P, Singh AK (2010) Molecular function prediction using neighborhood features. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)7:208-217.
[32]. Paul TK, Iba H (2009) Prediction of cancer class with majority voting genetic programming classifier using gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 6:353-367.