ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

A Review of Opinion Mining Approaches for Knowledge Discovery

Salina Adinarayana, Dr.E.Ilavarasan

Assoc Professor, Dept. of IT, Shri Vishnu Engineering College for Women, Bhimavaram, India

Professor, Dept. of CSE, Pondicherry Engineering College, Pondicherry, India

ABSTRACT: The knowledge discovery from data sources of social blogging is a challenging task. Since the data is not available in the specified format and the users are free to express their opinions in their desired style. Opinion mining is one of the latest and the most feasible approaches to promote business needs in the market. In this review paper, we tried to sum up all the needed technical and theoretical aspects for efficient implementation of data mining techniques for knowledge discovery from opinion mining datasets.

KEYWORDS: Knowledge Discovery, Data Mining, Classification, Opinion Mining, Social Blogging**.**

I. INTRODUCTION

Opinion mining is the process of discovering the sentiments of the users for a specific product or topic. The opinion mining data is available from the social networking websites, where users express their opinions regarding a product or a topic. The challenging task of opinion mining is about the data, which is available for sentiment analysis is of with different formats or styles. The data mining techniques are of great use for processing the data to prepare in the required format and then conduct the analysis for knowledge discovery. In this paper, we reviewed the recent approaches in the field of both classification and opinion mining for knowledge discovery. We have also provided a limited list of evaluation criteria's used in the field of opinion mining and the benchmark datasets used for sentiment analysis.

II. DATA MINING FOR KNOWLEDGE DISCOVERY

Data mining is the process of discovering hidden knowledge form the datasets. In data mining, the main broad approaches are classification and clustering. Classification is the process of predicting unknown instances using model build by training instances. In classification different models can be built for predicting the unknown instances of the data source. Some of the popular models are decision trees, regression, neural networks, support vector machine etc.

Data mining is the process of discovering hidden knowledge form the datasets. In data mining, the main broad approaches are classification and clustering. Classification is the process of predicting unknown instances using model build by training instances. In classification different models can be built for predicting the unknown instances of the data source. Some of the popular models are decision trees, regression, neural networks, support vector machine etc.

Some of the recent approaches proposed in the field of classification are given as follows: In [1] author proposed the Chi-FRBCS-Big Data algorithm, a linguistic fuzzy rule-based classification system that uses the MapReduce framework to learn and fuse rule bases. In [2] author presented the performance of ID3 classification and cascaded model with RBF network. In [3] author proposed a windowed regression over-sampling (WRO) method for oversampling of instances in the minority subset to change the class distribution through adding virtual samples. WRO not only reflects the additive effects but also reflects the multiplicative effect between samples.

In [4] author presented a review of existing solutions to the class-imbalance problem both at the data and algorithmic levels. In [5] author summarized a comprehensive study of different feature selection schemes in machine learning for the problem of mood classification in weblogs. A novel use of a feature set based on the affective norms for English words (ANEW) lexicon studied in psychology is also proposed. In [6] author presented a neural network-based finite impulse response extreme learning machine (FIR-ELM) for studying of medical datasets. In [7] author proposed a secure k-NN classifier over encrypted data in the cloud. The algorithm is used for solving the classification problem over encrypted data by protecting the confidentiality of data, privacy of user's input query and hides the data access patterns. Obviously, there are many other algorithms which are not included in this literature. A profound comparison of the above algorithms and many others can be gathered from the references list.

III. RECENT ADVANCES ON OPINION MINING

The field of opinion mining is one of the latest domains for analyzing the social blogging websites. One of the popular social blogging website is twitter. This section reviews the recent publication on twitter data analysis for mining knowledge hidden in the social websites.

Aamera Z.H. Khan et al., [8] have presented a new entity-level sentiment analysis method for Twitter in which lexicon based approach to perform entity-level sentiment analysis is adopted. Kun-Lin Liu et al., [9] have proposed a novel model, called emoticon smoothed language model (ESLAM), to handle manually and noisy labelled tweets data for training in an efficient way. They proposed to use manually labelled data in the training phase to build a good predictive model than to use noisy data for smoothing. Teu Terpstra et al., [10] have proposed the analyses of tweets through predefined (geo) graphical displays, message content filters (damage, casualties) and tweet type filters (e.g., retweets). The Important topics tracked in this work are 'early warning tweets', 'rumors' and the 'self-organization of disaster relief' on Twitter.

ChiranjibiSitaula et al., [11] have presented a novel approach for using stemmer technique on nepali language for improving the accuracy of analyzing the text content. OanaFrunza et al., [12] have proposed an approach for summarizing health care information by extracting the information from published papers by identifying semantic relations. M.Thangarasu et al., [13] have summarized about different stemming algorithms, specifically applicable on Indian languages. They had also stressed for the need of new proposals in the field for effective information retrieval. Andrea zieliski et al., [14] have conducted a study on problems of analyzing multilingual twitter feeds for emergency events using English as "lingua franca" and on under-resourced Mediterranean languages in endangered zones, particularly Turkey, Greece, and Romania Generally, as local civil protection authorities and the population are likely to respond in their native language. They also investigated ten earthquake events and defined four language-specific classifiers that can be used to detect earthquakes by filtering out irrelevant messages that do not relate to the event. Marion E. Hambrick et al., [15] have presented a study using content analysis to place 1,962 tweets by professional athletes into one of six categories: interactivity, diversion, information sharing, content, promotional, and fanship.

Saif Hassan et al., [16] have proposed two different sets of features to alleviate the data sparseness problem. One is the semantic feature set where they extract semantically hidden concepts from tweets and then incorporate them into classifier training through interpolation. Another is the sentiment-topic feature set where they extract latent topics and the associated topic sentiment from tweets, then augment the original feature space with these sentiment-topics. Apoorv Agarwal et al., [17] have proposed a POS-specific prior polarity features and tree kernel to obviate the need for tedious feature engineering in twitter data analysis. HaoWang et al., [18] have proposed a system for real time analysis of public sentiment toward presidential candidates in the 2012 U.S. election as expressed on Twitter, a micro blogging service.

Efthymios Kouloumpis et al., [19] have proposed an approach using linguistic features for detecting the sentiment of Twitter messages. They also used lexical resources as well as features to capture information about the informal and creative language used in micro blogging. Andranik Tumasjan et al., [20] have applied LIWC text analysis software and conducted a content analysis of over 100,000 messages containing a reference to either a political party or a politician. Daniel M. Romero et al., [21] have develop a formalization and methodology for studying link copying of directed closure process, and they provide evidence for its important role in the formation of links on Twitter. Alec Go et al., [22] have investigated different approaches using Naive Bayes, Maximum Entropy, and SVM classifiers for classifying sentiment of messages on micro-blogging services like Twitter. These are some of the recent contribution which used twitter for opinion mining.

IV. EVALUATION CRITERIA'S FOR OPINION MINING

The recent approaches in the opinion mining used the following evaluation criteria's:

To assess the classification results they count the number of true positive (TP), true negative (TN), false positive (FP) (actually negative, but classified as positive) and false negative (FN) (actually positive, but classified as negative) examples. It is now well known that error rate is not an appropriate evaluation criterion when there are unequal costs. In opinion mining AUC, Precision, F-measure, TP Rate and TN Rate are used as performance evaluation measures.

Apart from these simple metrics, it is possible to encounter several more complex evaluation measures that have been used in different practical domains. One of the most popular techniques for the evaluation of classifiers of unequal costs is the Receiver Operating Characteristic (ROC) curve, which is a tool for visualizing, organizing and selecting classifiers based on their tradeoffs between benefits (true positives) and costs (false positives).

The most commonly used empirical measure, accuracy does not distinguish between the number of correct labels of different classes, which in the framework of unequal costs may lead to erroneous conclusions. For example a classifier that obtains an accuracy of 90% in a dataset with a degree of imbalance 9:1, might not be accurate if it does not cover correctly any minority class instance.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{---------} \quad \text{(i)}$$

Because of this, instead of using accuracy, more correct metrics are considered. A quantitative representation of a ROC curve is the area under it, which is known as AUC. When only one run is available from a classifier, the AUC can be computed as the arithmetic mean (macro-average) of TP rate and TN rate:

The Area under Curve (AUC) measure is computed by,

$$AUC = \frac{1+TP_{RATE} - FP_{RATE}}{2} \quad \text{------------} \quad \text{(ii)}$$

Or

$$AUC = \frac{TP_{RATE} + TN_{RATE}}{2} \quad \text{------------} \quad \text{(iii)}$$

On the other hand, in several problems we are especially interested in obtaining high performance on only one class. For example, in the diagnosis of a rare disease, one of the most important things is to know how reliable a positive diagnosis is. For such problems, the precision (or purity) metric is often adopted, which can be defined as the percentage of examples that are correctly labeled as positive:

The Precision measure is computed by,

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{---------} \quad \text{(iv)}$$

The Recall measure is computed by,

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{----------} \quad \text{(v)}$$

DOI: 10.15680/IJIRCCE.2016. 0408052

The F-score value is computed by,

$$\text{F-Score} = \frac{2 \text{ X Precision X Recall}}{\text{Precision + Recall}} \quad \text{---------} \quad \text{(vi)}$$

To deal with unequal costs class, sensitivity (or recall) and specificity have usually been adopted to monitor the classification performance on each class separately. Note that sensitivity (also called true positive rate, TP rate) is the percentage of positive examples that are correctly classified, while specificity (also referred to as true negative rate, TN rate) is defined as the proportion of negative examples that are correctly classified:

The True Positive Rate measure is computed by,

$$TruePositi\ veRate = \frac{TP}{(TP)+(FN)} \quad \text{-----------} \quad \text{(vii)}$$

The True Negative Rate measure is computed by,

$$TrueNegati\ veRate = \frac{TN}{(TN)+(FP)} \quad \text{----------} \quad \text{(viii)}$$

V. BENCHMARK DATASETS USED IN OPINION MINING

Table 1 summarizes the benchmark datasets used in almost all the recent studies conducted on opinion mining. The details of the datasets are given in table 1. For each data set, the serial number, the name of the dataset and the category of the dataset is given. The details of the dataset are as follows:

Norton:
The Norton dataset belongs to the category of software. In this dataset, the different views such as positive, negative and neutral are presented. The example of reviews is given below ,
Example:

Norton products[-3]##But if I installed either one of these Norton products, neither works after installation?

##This package contains BOTH Norton Antivirus and Firewall.

Canon Power Shot SD 500:
The Canon power Shot SD 500 dataset belongs to category of digital camera. The user reviews for this product in the form of sentences is presented in the dataset. The example of reviews are given below,
Example:

SD500[+2]##We really enjoyed shooting with the Canon PowerShot SD500.
manual control[-1]##Some more manual control would have been nice.

Canon S100:
The Canon S 100 dataset belongs to category of digital camera. In this dataset, different reviews such as positive, neutral and negative are presented for analysis. The example of reviews are given below,

Example:

small[+1]##I want to start off saying that this camera is small for a reason.
flash[-2]##The flash is very weak.

Diaper Champ:

The Diaper Champ is house hold product used for children's by the parents. The parents expressed their views about the product. The sample of reviews is given below.

Example:

##We started with the Diaper Genie as most new parents do.

odor[-2]##My daughter is only 4 months old and we do notice an odor.

Table 1

Summary of Opinion Mining Datasets

S.no	Datasets	Category
1.	Norton	software
2.	Canon Power Shot SD 500	digital camera
3.	Canon S100	digital camera
4.	Diaper Champ	household
5.	Hitachi router	router
6.	Ipod	electronic
7.	Linksys Router	router
8.	Micro MP3	mp3 player
9.	Nokia 6600	cellular phone
10.	Speaker	electronic
11.	Computer	systems
12.	Wireless Router	router
13.	Canon G3 (Amazon customer reviews)	digital camera
14.	Nikon coolpix 4300 (Amazon customer reviews)	digital camera
15.	Nokia 6610 (Amazon customer reviews)	cellular phone
16.	Creative Labs Nomad Jukebox Zen Xtra 40GB (Amazon customer reviews)	mp3 player
17.	Apex AD2600 Progressive-scan DVD player (Amazon customer reviews)	dvd player
18.	Intel AMD (http://www.epinions.com)	processor
19.	Fuji FinePix Z1 (http://www.dcresource.com)	digital camera
20.	Fuji FinePix A210 (http://www.dcresource.com)	digital camera
21.	Fuji FinePix Z1 (http://www.dcresource.com)	digital camera
22.	Casio Exilim EX-Z750 (http://www.dcresource.com)	digital camera
23.	HP Pavilion dv4000 (PC World)	system
24.	Freestyle M7500 (Sys Technology)	system
25.	iPodnano	electronic
26.	Google vs Yahoo	discussion

27. Coke vs Pepsi	discussion
28. SFU Spanish Review Corpus (http://www.Ciao.es)	general reviews
29. Twitter	discussion
30. Google Talk	discussion

Hitachi router:
The Hitachi Router is one of the famous routers in the world market. The Hitachi router dataset provides the reviews of the product in terms of positive, negative and neutral. The examples of reviews are given below.
Example:
##Happy good working!!
router[+2]##Hitachi's M12V is a big beast of a router.

Ipod:
The Ipod dataset are of category electronic music device. This datasets presents views of different users on Ipod usage. The example of views is given below.

Example:
sound[+2]##SOUND QUALITY: The iPod's sound quality is pretty good.
games[-1]## Doesn't have a lot of good games on it.

Linksys Router:
The Linksys router dataset is of category router. This dataset provides the opinions of its user on the Linksys product. The sample of the reviews is given below.
Example:
router[+3][p] ##This linksys router does it for the RIGHT PRICE.
##provide no printed manual

Micro MP3:
The Micro Mp3 dataset is an opinion mining dataset of the category mp3 player. In this dataset the user views of positive, negative and neutral are presented. The example of the views from the dataset is given below.
Example:
player[+3][p] ##I've had this player for more than 4 months and I'm very happy with it.
microphone[+1] ##Record memos using the built in microphone.

These benchmark dataset provide the mainstream platform for most of the researchers for conducting sentiment analysis. The in depth details of the datasets can be gathered form the reference list. We hope that the review framework provided in this paper will be useful for the research community of opinion mining.

VI. CONCLUSION

In this paper, we reviewed the state of the art methodologies of data mining especially in classification and we also reviewed the recent proposal in the field of opinion mining. In our view, the recent methodologies in data mining and sentiment analysis should be combined to find the integrated solution to enhance the process of knowledge discovery in terms of sentiment analysis. We hope that, this review paper can serve as a good platform for the young researchers to better understand the problems for proposing better solutions for sentiment analysis in terms of knowledge discovery.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

REFERENCES

1. Sara del R, Victoria L´opez , Jos´e Manuel Ben´ıtez , Francisco HeCSera, "A MapReduce Approach to Address Big Data Classification Problems Based on the Fusion of Linguistic Fuzzy Rules", International Journal of Computational Intelligence Systems, Vol. 8, No. 3 (2015) 422-437.
2. Dharm Singh, Naveen Choudhary & JullySamota," Analysis of Data Mining Classification with Decision tree Technique:", Global Journal of Computer Science andTechnologySoftware& Data Engineering, Volume 13 Issue 13 Version 1.0 Year 2013.
3. Yong Hu, DongfaGuo, Zengwei Fan, Chen Dong, Qiuhong Huang, ShengkaiXie,Guifang Liu, Jing Tan, Boping Li, QiweiXie." An Improved Algorithm for Imbalanced Data and Small Sample Size Classification", Journal of Data Analysis and Information Processing, 2015, 3, 27-33.
4. Vaishali Ganganwar, "An overview of classification algorithms for imbalanced Datasets", International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 4, April 2012).
5. Thin Nguyen, DinhPhung, Brett Adams, Truyen Tran, and SvethaVenkatesh," Classification and Pattern Discovery of Mood inWeblogs, M.J. Zaki et al. (Eds.): PAKDD 2010, Part II, LNAI 6119, pp. 283–290, 2010.
6. Kevin Lee,ZhihongMan, Dianhui Wang, Zhenwei Cao," Classification of bioinformatics dataset using finite impulse response extreme learning machine for cancer diagnosis", Neural Comput&Applic, DOI 10.1007/s00521-012-0847-z.
7. Bharath K. Samanthula, Member, IEEE, YousefElmehdwi, and Wei Jiang, Member, IEEE,"k-Nearest Neighbor Classification over Semantically Secure Encrypted Relational Data, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 5, MAY 2015.
8. AameraZ.H.Khan, Mohammad Atique, V. M. Thakare, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis" National Conference on "Advanced Technologies in Computing and Networking"-ATCON-2015, Special Issue of International Journal of Electronics, Communication & Soft Computing Science and Engineering, ISSN: 2277-9477
9. Kun-Lin Liu,Wu-Jun Li, MinyiGuo," Emoticon Smoothed Language Models for Twitter Sentiment Analysis", Association for the Advancement of Artificial Intelligence (www.aaai.org), 2012.
10. TeuTerpstra, A.de Vries, R.stronkman, G.L.paradies," Towards a real time Twitter analysis during crises for operational crisis management",Proceedings of the 9th International ISCRAM Conference – Vancouver, Canada, April 2012 L. Rothkrantz, J. Ristvej and Z. Franco, eds.
11. ChiranjibiSitaula,"A Hybrid Algorithm for Stemming of Nepali Text, Intelligent Information Management, 2013, 5, 136-139 doi:10.4236/iim.2013.54014 Published Online July 2013 (http://www.scirp.org/journal/iim)
12. Oana Frunza, Diana Inkpen, Thomas Tran, "A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 6, JUNE 2011
13. M.Thangarasu, Dr.R.Manavalan,"A Literature Review: Stemming Algorithms forIndian Languages, International Journal of Computer Trends and Technology (IJCTT) – volume 4 Issue 8–August 2013
14. Andrea zieliski, Ulrich Bugel, "Multilingual Analysis of Twitter News in Support of Mass Emergency Events",Proceedings of the 9th International ISCRAM Conference – Vancouver, Canada, April 2012 L. Rothkrantz, J. Ristvej and Z. Franco, eds.
15. Marion E. Hambrick, Jason M. Simmons, Greg P. Greenhalgh, T. Christopher Greenwell," International Journal of Sport Communication, 2010, 3, 454-471© 2010 Human Kinetics, Inc.
16. Saif, Hassan; He, Yulan and Alani, Harith (2012). Alleviating data sparsity for Twitter sentiment analysis. In: 2nd Workshop on Making Sense of Micro posts (#MSM2012): Big things come in small packages at the 21st International Conference on theWorld Wide Web (WWW'12), 16 April 2012, Lyon, France, CEUR Workshop Proceedings (CEUR-WS.org), pp. 2–9.
17. ApoorvAgarwal, BoyiXie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau," Sentiment Analysis of Twitter Data", Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 30–38, Portland, Oregon, 23 June 2011. c 2011 Association for Computational Linguistics.
18. Hao Wang, Dogan Can, Abe Kazemzadeh," A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle", Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 115–120, Jeju, Republic of Korea, 8-14 July 2012. c 2012 Association for Computational Linguistics.
19. EfthymiosKouloumpis, TheresaWilson, Johanna Moore," Twitter Sentiment Analysis: The Good the Bad and the OMG!", Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media,
20. AndranikTumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe," Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment", Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.
21. Daniel M. Romero, Jon Kleinberg," The Directed Closure Process in Hybrid Social-Information Networks,with an Analysis of Link Formation on Twitter", Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.
22. Go, A.; Bhayani, R. & Huang, L. (2009), 'Twitter Sentiment Classification using Distant Supervision', Processing, 1--6.