INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 3 (2017) PAGES 2469 - 2475
RECEIVED : 10.05.2017 PUBLISHED : 17.06.2017
June 17, 2017

DECISION TREE BASED METHOD FOR PREDICTING QUESTION SUBJECTIVITY IN SOCIAL QUESTION AND ANSWERING

[1]Prajisha.C, [2]Mr.N.Ashok Kumar,
[1]PG scholar, Department of Computer Science and Engineering, Maharaja Engineering College , avinashi,
[2]Guide, Assistant Professor, Department of Computer Science and Engineering, Maharaja Engineering College , avinashi.

ABSTRACT

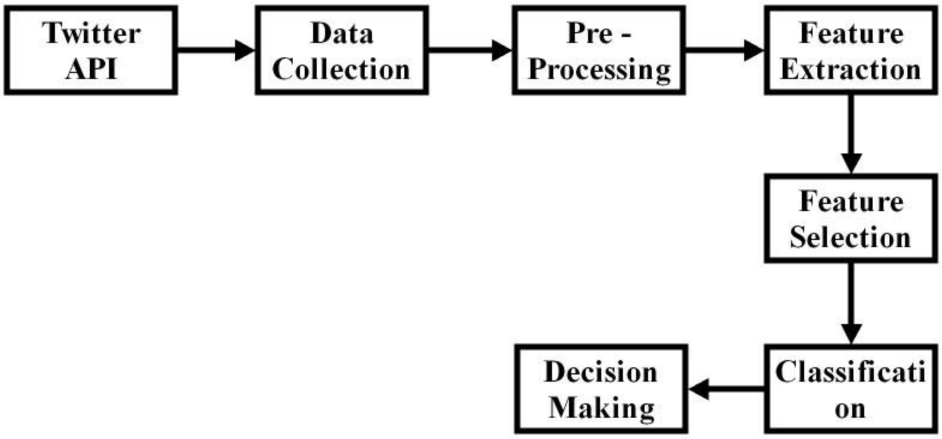
The rise of long range interpersonal communication destinations (SNSs, for example, Face book and Twitter, has made the correspondence among people more various and advantageous. Other than utilizing those social stages for relationship upkeep, many individuals likewise see SNSs as important data sources and participate in what has been alluded to as social question and replying (social Q&A) Contrasted and the common web index administrations, for example, Google and Bing, social Q&A gives individuals a more straightforward and simpler approach to express their data needs, as people can openly communicate their demand for help in normal dialects to all companions or adherents on the web, and to get more customized and dependable responses. Using basic components extricated from the question message, this technique can consequently recognize the subjectivity introduction of an examiner's goal. Via consequently recognizing subjective inquiries from the goal ones, one could at last form address steering frameworks that can guide a question to its potential answerers as per its fundamental plan. For example, given a subjective question, we could course it to some person who knows the examiner well to give more customized reactions. Be that as it may, for a goal address, we could find specialists inside a specific space or could consequently answer another question utilizing the chronicled question–answer sets.

Keywords: Social Networks, , Web index, Component Extraction, Subjectivity, Potential answers.

1. RELATED WORKS

Li revealed that there were about 11% of general tweets containing questions and 6% of tweets having information needs[1]. Going one step further, Efronand Winget analyzed 100 question tweets on Twitter and proposed a taxonomy of questions asked on micro blogging platforms. Morris et al. manually labelled a set of questions posted on social networking platforms and identified eight question types in social Q&A, including recommendation, opinion, factual knowledge rhetorical, invitation, favour, social connection, and offer [1]. Social networking services provide a source of information that is complementary to that provided by search engines, the former provides information that is highly tailored to an individual and comes from a highly trusted source, while the latter provides objective data from a variety of sources on a variety of topics [2]. To better understand social network Q&A exchanges we conducted a survey of 624 people using social networking services like Face book and Twitter. Our survey covered topics such as the prevalence of asking and answering questions via status-message updates, the types and topics of the questions asked, the speed and quality of the answers received, and the motivations people have for asking and answering questions on social networks. Our analysis also explores the influence of properties of the question and demographics of the asker on response speed and quality [2]. There are only a few papers that touch on the problem of automatic question classification based on machine learning techniques. Li et al. proposed a cascade approach, which first detected interrogative tweets and then questions revealing real information needs (referred to as Tweets in their paper). They relied on both rule-based and learning-based approaches for interrogative tweets detection and some Twitter-specific features, such as re tweet, mentioned to extract tweets [3]. As a result of their analysis, they claimed that conversational questions typically have much lower potential archival value than the informational ones. Kim et al. classified questions from Yahoo! Answers into four categories: information, suggestion, opinion, and other. They pointed out that the criteria of selecting best answer differed across categories. Pal et al. introduced the concept of question temporality based on when the answers provided on the questions would expire. They labelled questions into five categories, with permanent, long, medium, short, and other temporal durations [4]. . Questions that convey information needs are extracted from a collection of billions of micro blogs (i.e., tweets). This is achieved by an automatic text classier that distinguishes real questions (i.e., tweets conveying real information needs) from tweets with question marks. With this dataset, we are able to present a comprehensive description of the information needs with both the perspectives of content analysis and trend analysis [5].

2. THE MODEL



The question classification system is an important part of most of the data mining techniques. The proposed method uses multi-label decision tree classification algorithm and Naïve Bayes classification algorithm for classifying questions in SNSs. These algorithms are more efficient than binary classification algorithms with respect to noise reduction. The proposed method uses Twitter API to retrieve tweets containing questions. Initially, the retrieved tweets are saved into a csv file. Preprocessing stage removes the tweets does not containing questions, Re-tweets, noises, etc. After preprocessing extract important features using feature extraction methods such as count Vector and Tf-Idf Transformer. Select top features from extracted features using Chi square feature selection methods. The selected features and labels are given to the Multi-label decision tree classifier for identifying different types of questions. The detailed description of the proposed method is described in the following sections.

3. THE ARCHITECTURE

In proposed design a few tweets are gathered in light of hash tags#engineeringProblem,

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 3 (2017) PAGES 2469 - 2475
RECEIVED : 10.05.2017 PUBLISHED : 17.06.2017 June 17, 2017

#nerdstatus, and tweets . These assistance in depicting the procedure to find the pertinent inquiries (a Twitter hash tag is a word starting with a # sign, used to accentuate or tag a point). In the beneath figure the width of dark bolt speaks to information volume more extensive shows more information volume. Light dark bolts speak to information examination, calculation, and result stream. The stream can be compressed in the accompanying strides:

- Data is collected from social media content.
- A detailed pre-processing is done.
- Extract best features from training data set.
- Select best features from Extracted features.
- Questions are categorized and a multi-label classifier is proposed which can be implemented by decision tree classification algorithm and Naïve Byes algorithm.
- The result could help users identify the subjectivity of questions.

4. IMPLEMENTATION

Naive Bayes Multi label Classifier

Transformation of the multi-label classification problem into multiple single-label classification problems is one of the popular ways to implement the multi-label classifier. One-versus-all or Binary Relevance is one of the transformation methods which consists of assuming the independence among categories, and train a binary classifier for each category. All kinds of binary classifier can be transformed to multi-label classifier using the one- versus-all heuristic. The following are the basic procedures of the Naive Bayes multi-label classifier.

Assume there are a total number of N words in the training document (for our situation, every tweet is a document) W = {w1, w2, ...,wN } , and a total number of L categories C = {c1, c2, ..., cL}. If a word wn appears in a category c for mwnc times, and appear in categories other than c for mwnct times, then based on the Maximum Likelihood Estimation, the probability of this word in a specific category c is.

$$\mathrm{p}(w_n, \mathrm{c}) = m_{w_n}c / \textstyle\sum_{n=1}^{N} m_{w_n}c$$

Similarly, the probability of this word in categories other than c is

$$\mathrm{P}(w_{n/c^t}) = \frac{m_{w_n c^t}}{\sum_{n=1}^{N} m_{w_n} c^t}$$

Assume there are a total number of M documents in the training data set, and C of them are in category c. At that point the probability of category c is

$$\mathrm{P(c)} = \frac{C}{m^t}$$

and the probability of other than categories c is

$$\mathrm{p(c)} = \frac{M-C}{m^t}$$

For a document di in the testing data set, there are K words Wdi ={wi1, wi2, ..., wiK}, and Wdi is a subset of W. The objective is to classify this doc- ument into category c or not c. We assume independence among each word in this document, and any word wik conditioned on c or c' follows multinomial distribution. Therefore, according to Bayes Theorem, the probability that di belongs to category c is

$$P(c/d_i) = \frac{P\left(\frac{d_i}{c}\right).P(c)}{P(d_i)} \alpha \prod_{k=1}^{k} p(w_{ik}/c).p(c)$$

and the probability that di belongs to categories other than c is

$$P(c^t/d_i) = \frac{p(d_{i/}c^t)}{P(d_i)} \alpha \prod_{k=1}^{k} p(w_{ik}/c^t).p(c)$$

Because p(c/d_i) + p(c^t/d_i) = 1, normalize the latter two items which are propor- tionalto p(c/d_i) and p(c^t/d_i) to get the real values of p(c/d_i). If p(c/d_i) is larger than the probability threshold T, then di belongs to category c, otherwise, di does belong to category c. Then repeat this procedure for each category.

5. CLASSIFICATION RESULTS

For multi-label classification there are usually two types of evaluation measures example-based measures and label-based measures. Example- based measures are calculated on each document (e.g. each tweet is a document, and also called an example here) and then averaged over all documents in the dataset, whereas label-based measures are calculated based on each label (category) and then averaged over all labels (categories). In each of the one-versus-Rest binary classification step the performance measures are,

Accuracy: Accuracy is simply the ratio of correctly predicted observations.

Accuracy a=$\frac{tp+tn}{tp+tn+fp+fn}$

Precision: Precision is the ratio of correct positive observations.

Precision p=$\frac{tp}{tp+fp}$

Recall: Recall is also known as sensitivity or true positive rate. Its the ratio of correctly predicted positive events.

Recall r=$\frac{tp}{tp+fn}$

F1-Score: The F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into ac- count. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if have an uneven class distribution. It works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

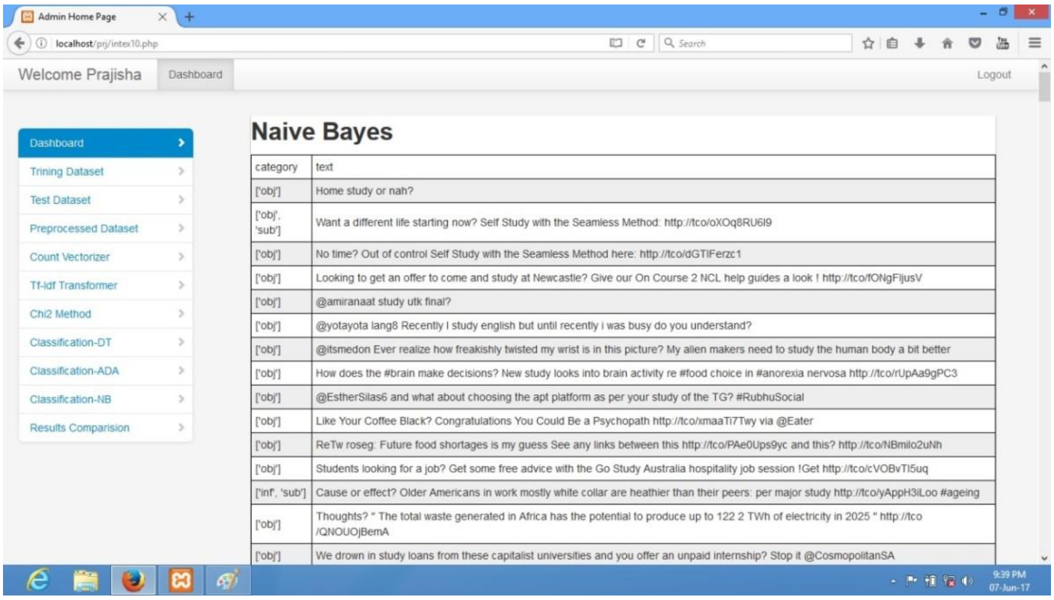
F1 Score=$\frac{2tp}{2tp+fp+fn}$

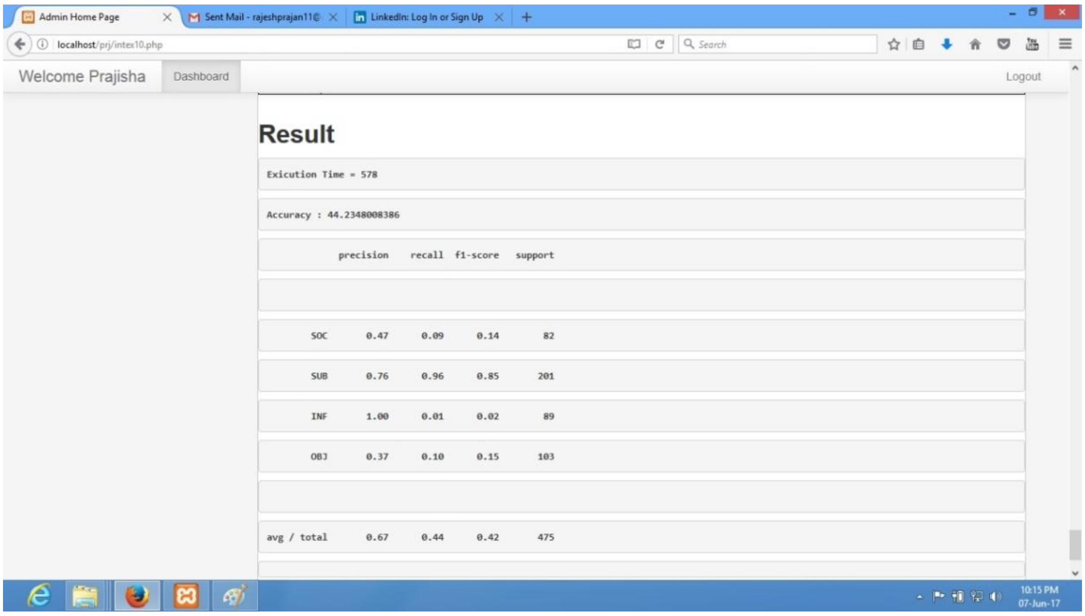
Where, TP is True positive, FP is False positive, TN is True Negative, FN is False Negative. True positive rate measures the proportion of posi- tives that are correctly identified as positive. False positive rate measures the proportion of positives that are incorrectly identified as positive. True negative rate measures the proportion of negatives that are correctly iden- tified as negative. False negative rate measures the proportion of negatives that are incorrectly identified as negative.

CATEGORIES	ACCURACY	PRECISION	RECALL	F1-SCORE
DECISION TREE	43.56	0.52	0.43	0.43
NAÏVE BAYES	44.23	0.67	0.42	0.42

6. RESULTS

Transformation of the multi-label classification problem into multiple single label classification problems is one of the best way to implement the multi-label classifier. One-versus-all or Binary Relevance is one of the transformation methods which consists of assuming the independence among categories, and train a binary classifier for each category. All kinds of binary classifier can be transformed to multi-label classifier using the one- versus-all heuristic. The following are the basic procedures of the Decision tree multi-label classifier.





CONCLUSION

There are numerous constraints for the manual subjective investigation and substantial scale computational examination of client created printed content. Machine learning based classifiers help the analysts in learning analytics. This prescient model on question subjectivity empowers programmed identification of subjective and target data looking for inquiries posted on Twitter and can be utilized to encourage future reviews on vast scales. This investigation comes about enable the specialists to comprehend the particular goals behind subjective and target addresses and to construct relating apparatuses or frameworks to better upgrade the coordinated effort among people in supporting social Q&A exercises. For example, we imagine that given the study way of subjective inquiries and more peculiar's interests in noting them, one could build up a calculation to course those subjective inquiries to suitable respondents in light of their areas and past encounters. Conversely, considering the factorial nature and brief term of target inquiries, they could be steered to either web search tools or people with comparable ability or accessibility.

REFFERENCE

1. AgichteinE., GarciaV.,LiB., LiuY.,Ram A.,(2008) "Exploring question subjectivity prediction in community QA," in Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr, pp. 735–736.
2. Bernard J. Jansen.,ZheLiu (2016)"Understanding and Predicting Question Subjectivity in Social Question and Answering"IEEE Transactions On Computational Social Systems, VOL. 3, NO. 1
3. MeiQ.,Zhao Z.,(2009) "Questions about questions: An empirical analysis of information needs on Twitter," in Proc. 22nd Int. Conf. World Wide Web, , pp. 1545–1556.
4. MorrisM. R.,PanovichK.,Teevan J., (2010) "What do people ask their social networks, and why?: A survey study of status message q&abehavior," in Proc. SIGCHI Conf. Human Factors Comput. Syst.
5. Wilson T..,(2005) "OpinionFinder: A system for subjectivity analysis," in Proc. HLT/EMNLP Interact. Demonstrations , pp. 34–35.
6. A multisession-based multidimensional model. M. Body, M. Miquel, Y. B´edard, and A. Tchounikine. A multidimensional and multi version structure for OLAP applications. In DOLAP '02: Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP, pages 1–6, New York, NY, USA,2002. ACM.
7. Transaction Management for a Main-Memory Database. P. Burte, B. Aleman-meza, D. B. Weatherly, R. Wu, S. Professor, and J. A. Miller. Transaction Management for a Main-Memory Database. The 38th Annual South eastern ACM Conference, Athens, Georgia, pages 263–268, January 2001.
8. Discovering business intelligence from online product reviews: A rule-induction framework. W. Chung and H. Chen. Web-Based Business Intelligence Systems: A Review and Case Studies. In G. Adomavicius and A. Gupta, editors, Business Computing, volume 3, chapter 14, pages 373–396. Emerald Group Publishing, 2009.
9. Crowd sourcing Predictors of Behavioural Outcomes. Josh C. Bongard, Member, IEEE, Paul D. H. Hines, Member, IEEE, Dylan Conger, Peter Hurd, and Zhenyu Lu. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING YEAR 2013.
10. Feature Selection Based on Class-Dependent Densities for High-Dimensional Binary Data. Kashif Javed, Haroon A. Babri, and Mehreen Saeed. Ieee transactions on knowledge and data engineering, vol. 24, no. 3, march 2012.
11. Techniques, Process, and Enterprise Solutions of Business Intelligence. Li Zeng, Lida Xu, Zhongzhi Shi, Maoguang Wang, and Wenjuan Wu. 2006 IEEE Conference on Systems, Man, and Cybernetics October 8-11, 2006, Taipei, Taiwan.

12. Support vector machine with adaptive parameters in financial time series forecasting. Cao, L.J. ; Dept. of Mech. Eng., Nat. Univ. of Singapore, Singapore ; Tay, F.E.H. EEE Transactions on, On page(s): 1167 - 1178 Volume: 19, Issue: 7, July 2008
13. Financial time series modelling with discounted least squares back-propagation. A. N. Refenes, Y. Bentz, D. W. Bunn, A. N. Burgess, and A. D. Zapranis, "Financial time series modeling with discounted least squares back-propagation", Neuro computing, vol. 14, pp.123 -138 1997. Stock Market Value Prediction Using Neural Networks.
14. T. Kimoto, K. Asakawa, M. Yoda, and M. Takeoka, "Stock market prediction system with modular neural networks", Neural Networks in Finance and Investing, pp.343 -357 1993.
15. Application of a Case Base Reasoning Based Support Vector Machine for Financial Time Series Data Forecasting. Pei-Chann Chang, Chi-Yang Tsai, Chiung-Hua Huang, Chin-Yuan Fan 5th International Conference on Intelligent Computing, ICIC 2009 Ulsan, South Korea, September 16-19, 2009 Proceedings.