ISSN: 2321-7782 (Online)
Volume 2, Issue 10, October 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Mining Social Networking Data for Classification Using Reptree

Dr. B. Srinivasan[1]
Associate Professor of Comp. Science
PG & Research Dept. of Comp. Science
Gobi Arts & Science College (Autonomous)
Gobichettipalayam – 638453

P.Mekala[2]
M.Phil Research Scholar
PG & Research Dept. of Comp. Science
Gobi Arts & Science College (Autonomous)
Gobichettipalayam – 638453

Abstract: ***On various social media sites, students discuss and share their everyday encounters in an informal and casual manner. Analyzing such data, however, can be challenging. The complexity of students' experiences reflected from social media content requires human interpretation. However, the growing scale of data demands automatic data analysis techniques. This work focus on to demonstrate a workflow of social media data sense-making for educational purposes, integrating both qualitative analysis and large-scale data mining techniques and to explore engineering students' informal conversations on social media, in order to understand issues and problems students encounter in their learning experiences and this also gives the protection of students. In existing system the Naivy Bayes classifier is used to classify the twitter dataset thus increases the mean squared error. The Rep Tree classification technique will prove the proposed works efficiency.***

Keywords: Social media sites, social media data, classification technique, Naivy Bayes, Rep Tree

I. INTRODUCTION

Social Networking Internet services are changing the way to communicate with others, entertain and actually live. Social Networking is one of the primary reasons that many people have become avid Internet users; people who until the emergence of social networks could not find interests in the web. This is a very robust indicator of what is really happening online. The rapid growth in popularity of social networks has enabled large numbers of users to communicate, create and share content, give and receive recommendations, and, at the same time, it opened new challenging problems. The unbounded growth of content and users pushes the Internet technologies to its limits and demands for new solutions.

The emerging field of learning analytics and educational data mining has focused on analyzing structured data obtained from course management systems (CMS), classroom technology usage, or controlled online learning environments to inform educational decision-making [1]. However, to the best of our knowledge, there is no research found to directly mine and analyze student- posted content from uncontrolled spaces on the social web with the clear goal of understanding students' learning experiences.

A. Social Media Data

Social media is the social interaction among people in which they create, share or exchange information and ideas in virtual communities and networks. Social media is defined as a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0 and that allow the creation and exchanges of user-generated content. Social media is conglomerate of different types of social media sites including traditional media such as newspaper, radio, and television and non-traditional media such as Facebook, Twitter, etc.

The research goals of this study are 1) to demonstrate a workflow of social media data sense-making for educational purposes, integrating both qualitative analysis and large-scale data mining techniques to explore engineering students' informal conversations on Twitter, in order to understand issues and problems students encounter in their learning experiences.

Basically Reduced Error Pruning Tree ("REPT") is fast decision tree learning and it builds a decision tree based on the information gain or reducing the variance. REP Tree is a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning. It only sorts values for numeric attributes once. Missing values are dealt with using C4.5's method of using fractional instances. The example of REP Tree algorithm is applied on UCI repository and the confusion matrix is generated for class gender having six possible values.

II. Literature Review

Web (data) mining is one of the intelligent computing techniques in the context of Web data management. In general, Web mining is the means of utilizing data mining methods to induce and extract useful information from Web data information. Web mining research has attracted a variety of academics and engineers from database management, information retrieval, artificial intelligence research areas, especially from data mining, knowledge discovery, and machine learning etc. Basically, Web mining could be classified into three categories based on the mining goals, which determine the part of Web to be mined: Web content mining, Web structure mining, and Web usage mining [3]. Web content mining tries to discover valuable information from Web contents (i.e. Web documents). Generally, Web content is mainly referred to textual objects, thus, it is also alternatively termed as text mining sometimes [2]. Web structure mining involves in modeling Web sites in terms of linking structures. The mutual linkage information obtained could, in turn, be used to construct Web page communities or find relevant pages based on the similarity or relevance between two Web pages. A successful application addressing this topic is finding relevant Web pages through linkage analysis [3].

Rost et al [5] argue that in large scale social media data analysis, faulty assumptions are likely to arise if automatic algorithms are used without taking a qualitative look at the data. To concur with this argument, as found no appropriate unsupervised algorithms could reveal in-depth meanings in our data. For example, LDA (Latent Dirichlet Allocation) is a popular topic modeling algorithm that can detect general topics from very large scale data. LDA has only produced meaningless word groups from our data with a lot of overlapping words across different topics.

Web applications are increasing at an enormous speed and its users are increasing at exponential speed. The evolutionary changes in technology have made it possible to capture the users' essence and interactions with web applications through web server log file. Web log file is saved as text (.txt) file. Due to large amount of "irrelevant information" in the web log, the original log file cannot be directly used in the web usage mining (WUM) procedure. Therefore the preprocessing of web log file becomes imperative. The proper analysis of web log file is beneficial to manage the web sites effectively for administrative and users' prospective. Web log preprocessing is initial necessary step to improve the quality and efficiency of the later steps of WUM. There are number of techniques available at preprocessing level of WUM. Different techniques are applied at preprocessing level such as data cleaning, data filtering, and data integration. In this research, survey the preprocessing techniques to identify the issues and how WUM preprocessing can be improved for pattern mining and analysis [4].

III. Proposed Methodology

Because social media content like tweets contain a large amount of informal language, sarcasm, acronyms, and misspellings, meaning is often ambiguous and subject to human interpretation.

There were no pre-defined categories of the data, so needed to explore what students were saying in the tweets. So need to conduct an inductive content analysis on the #engineering Problems dataset. Inductive content analysis is one popular qualitative research method for manually analyzing text content. Three researchers collaborated on the content analysis process. Analysis was to identify what are the major worries, concerns, and issues that engineering students encounter in their study and

life. Researcher A read a random sample of 2,000 tweets from the 19,799 unique #engineering Problems tweets, and developed 13 initial categories including: curriculum problems, heavy study load, study difficulties, imbalanced life, future and carrier worries, lack of gender diversity, sleep problems, stress, lack of motivation, physical health problems, nerdy culture, identity crisis, and others. These were developed to identify as many issues as possible, without accounting for their relative significances. Researcher A wrote detailed descriptions and gave examples for each category and sent the codebook and the 2,000-tweet sample to researchers B and C for review. Then, the three researchers discussed and collapsed the initial categories into five prominent themes, because they were themes with relatively large number of tweets. The five prominent themes are: *heavy study load, lack of social engagement, negative emotion, sleep problems, and diversity issues*. Each theme reflects one issue or problem that engineering students encounter in their learning experiences.

In this research, found that many tweets could belong to more than one category. For example, "This could very well turn into an all nighter...f*** you lab report #nosleep" falls into heavy study load, lack of sleep, and negative emotion at the same time. "Why am I not in business school?? Hate being in Engineering school. Too much stuff. Way too complicated. No fun" falls into heavy study load, and negative emotion at the same time. So one tweet can be labeled with multiple categories. This is a multi-label classification as opposed to a single-label classification where each tweet can only be labeled with one category. The categories one tweet belongs to are called this tweet's labels or label set.

Inter-rater Agreement

Statistical measures such as Cohen's Kappa, Scott's P_i, Fleiss Kappa, and Krippendorf's Alpha are widely used to report agreement among raters (researchers) in content analysis literature. However, these measures can only be used for data that belong to mutually exclusive categories (single-label classification). Because dealing with a multi-label classification problem with non-mutually exclusive categories, these measurements were inapplicable for our study. Therefore, F_1 measure is used which is the harmonic mean between two sets of data. F_1 score is 1 when the two sets of data are exactly the same, and is 0 if the two sets of data are completely different. It represents how close two label sets are assigned to one tweet by two researchers.

Then calculated the F1 scores between the label sets given by any two researchers to a tweet, and then averaged over all the tweets to represent the agreement. Assume there are a total number of N tweets categorized by two researchers. For the i-th tweet, x_{1i} represents the number of labels given to this tweet by researcher A, x_{2i} represents the number of labels given to this tweet by researcher B, and s_i represents the number of labels that are common between researcher A and researcher B (the agreed number of labels). Let $p_{1i} = s_i/x_{1i}$, and $p_{2i} = s_i/x_{2i}$, then

$$F_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{2\, P_{1i} \,.\, P_{2i}}{P_{1i} + P_{2i}}$$

In this study, after decided on the six categories (heavy study load, lack of social engagement, negative emotion, sleep problems, diversity issues, and others), took a random sample of 500 tweets, and categorized them separately. If a tweet does not convey any of the five prominent problems, it is categorized as "others". A tweet in "others" can be an engineering student problem other than the five prominent ones, or a noisy tweet that does not have clear meaning. Unlike the five prominent themes, "others" is an exclusive category.

The F1 scores between any of the two researchers were F_{1AB}=0.7972, F_{1BC}=0.7972 and F_{1AC}=0.8179. The three researchers then discussed tweets. The first 500 tweets were discarded, and analyzed another random sample of 500 tweets. The F1 scores then increased to F_{1AB}= 0.8104, F_{1BC}=0.8011, and F_{1AC}=0.8252 respectively. For the second 500-tweet sample, we only used the commonly agreed labels by the three researchers for each tweet in our later computation stage. If there was no intersection among the three researchers' label sets for a certain tweet, the tweet was discarded. Out of the 500 tweets, 405 were kept and

used in model training and testing. Researcher a then finished analyzing another random sample of 2,380 tweets. Plus the 405 tweets, there were a total of 2,785 labeled tweets used for model training and testing.

A. REP Tree

Basically Reduced Error Pruning Tree ("REPT") is fast decision tree learning and it builds a decision tree based on the information gain or reducing the variance. REP Tree is a fast decision tree learner which builds a decision/regression tree using information gain as the splitting criterion, and prunes it using reduced error pruning. It only sorts values for numeric attributes once. Missing values are dealt with using C4.5's method of using fractional instances. The example of REP Tree algorithm is applied on UCI repository and the confusion matrix is generated for class gender having six possible values.

In regression tree RT [E; Y], E is the leaf of the tree where the tree ends and Y is the response variable. Finding a binary question which gives the maximum information about the Y should be identified and the process should repeat for all levels of the tree. Here Y is considered to be the spam branch of the tree. The leaf E should give the maximum information about this branch that better discriminates the spam and genuine sites. In each children node the process should be repeated in greedy manner. And finally it yields a tree with maximum information gain of spam websites. Since the algorithm is recursive it requires stopping criteria. It is a threshold here. The sum of squared errors for a tree RT is defined as,

$$S = \sum_{Eeleaves(RT)} \sum_{i \in E} (Y_i - N_T)^2$$

Where, N_T is defined as

$$N_T = \frac{1}{P_c} \sum_{i \epsilon T} Y_i$$

The above equation is for N_T . And then formula becomes,

$$S = \sum_{Eeleaves(RT)} p_c V_c$$

where, V_c is considered to be the within variance and p_c is the class prediction

IV. IMPLEMENTATION

Commonly used measures to evaluate the performance of classification models include accuracy, precision, recall, and the harmonic average between precision and recall – the F1 score. For multi-label classification, the situation is slightly more complicated, because each document gets assigned multiple labels. Among these labels, some may be correct, and others may be incorrect. Therefore, 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).

Accuracy

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}$$

It is calculated by the above equation, where True positive is correctly identified, false positive is incorrectly identified, True negative is correctly rejected and false negative is incorrectly rejected.

A. Classification results

From the inductive content analysis stage, we had a total of 2,785 engineering Problems tweets annotated with 6 categories. We used 70% of the 2,785 tweets for training (1,950 tweets), and 30% for testing (835 tweets). 85.5% (532/622) of words occurred more than once in the testing sets were found in the training data set. Table 2 shows the 6 evaluation measures at each probability threshold values from 0 to 1 with a segment of 0.1. We assigned the one category with the largest probability value to the document when there was no category with a positive probability value larger than T. So when the probability threshold was 1, it was equivalent to outputting the largest possible one category for all the tweets.

With five multi-label categories and one "others" category, there are $(2^5-1)+1=32$ possible label sets for a tweet. Table 2 and Table 3 provide all the evaluation measures under random guessing. The random guessing program first guessed whether a tweet belongs to "others" based on the proportion this category takes in the training dataset. If this tweet did not belong to "others", it then proceeded to guess whether it fell into the rest of the categories also based the proportion each category takes in the rest categories. We repeated the random guessing program 100 times, and obtained the average measures.

TABLE I
Evaluation Measures with SVM Classifier under Different Probability Thresholds

Probability Threshold	example-based accuracy	example-based precision	micro-averaged F1	macro-averaged F1
0	0.1720	0.1720	0.2940	0.2550
0.2	0.6620	0.6670	0.6840	0.6795
0.5	0.7018	0.7090	0.7050	0.6115
0.8	0.7060	0.7156	0.7050	0.6005
1	0.7086	0.7198	0.7076	0.6026
Rand	0.0414	0.0416	0.0392	0.0180

From Table 1, see that when the probability threshold value is 0.4, the performance is generally better than under other threshold values.

TABLE II
Evaluation Measures

example-based accuracy	example-based precision	example-based recall	example-based F1
0.5518	0.6058	0.5830	0.5800

B. Experimental results

The total of 65535 tweets is used for classification. 66% of the 65535 tweets are used for training (43253 tweets) and 34% of the 65535 tweets for testing (22282 tweets). For this dataset, the result of the Naivy Bayes classifier increases the mean squared error (0.1913). The following table shows the evaluation measures with Naivy Bayes Classifier.

TABLE III
Evaluation Measures with Naivy Bayes Classifier

TP Rate	FP Rate	Precision	Recall	F-Measure
93.9	8.3	94.7	93.9	94.2

The same dataset is applied to Rep tree classifier. It reduces the mean squared error (0.1193) and also it increases the classification accuracy.

TABLE IV
Evaluation Measures with Naivy Bayes Classifier

TP Rate	FP Rate	Precision	Recall	F-Measure
98.1	8.4	98	98.1	98

V. CONCLUSION

The interpretation of dielectric behavior of saliva in terms of its molecular structure is a scientific objective. The study is useful to understand the physical phenomenon that occurs in dielectric that is placed in alternating field and to find the parameters of the dielectric, which quantitatively determine their electrical propertieskk.

In this study, through a qualitative content analysis, found that engineering students are largely struggling with the heavy study load, and are not able to manage it successfully. Heavy study load leads too many consequences including lack of social engagement, sleep problems, and other psychological and physical health problems.

Many students feel engineering is boring and hard, which leads to lack of motivation to study and negative emotions. Diversity issues also reveal culture conflicts and culture stereotypes existing among engineering students. Building on top of the qualitative insights, now implemented and evaluated a multi-label classifier to detect engineering student problems from University.

This detector can be applied as a monitoring mechanism to identify at-risk students at a specific university in the long run without repeating the manual work frequently. This is beneficial to researchers in learning analytics, educational data mining, and learning technologies. It provides a workflow for analyzing social media data for educational purposes that overcomes the major limitations of both manual qualitative analysis and large scale computational analysis of user-generated textual content. And study can inform educational administrators, practitioners and other relevant decision makers to gain further understanding of engineering students' college experiences.

References

1. S. Cetintas, L. Si, H. Aagard, K. Bowen, and M. Cordova- Sanchez, "Microblogging in Classroom: Classifying Students' Relevant and Irrelevant Questions in a Microblogging- Supported Classroom," Learning Technologies, IEEE Transactions on, vol. 4, no. 4, pp. 292–300, 2011.
2. S. Chakrabarti. Data mining for hypertext: a tutorial survey. SIGKDD Explor. Newsl., 1(2):1–11, 2000.
3. J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explor. Newsl., 1(2):12–23, 2000.
4. Hussain, T.; Asghar, S.; Masood, N., "Web usage mining: A survey on preprocessing of web log file," Information and Emerging Technologies (ICIET), 2010 International Conference on , vol., no., pp.1,6, 14-16 June 2010.
5. J. Yang and S. Counts, "Predicting the speed, scale, and range of information diffusion in twitter," Proc. ICWSM, 2010.

