RESEARCH ARTICLE OPEN ACCESS

A Survey on Analysing Students Learning Experiences by Extracting Data from Social Media (Social Forums)

Aditi Verma[1], Rachana Agarwal[2], Sameer Bardia[3], Simran Shaikh[4]

[1,2,3,4](Computer Engineering Department, Dr. D. Y. Patil Institute of Engineering & Technology, Pimpri, Savitribai Phule Pune University,Pune)

Abstract:

There are various online networking sites such as Facebook, twitter where students casually discuss their educational experiences, their opinions, emotions, and concerns about the learning process. Information from such open environment can give valuable knowledge for opinions, emotions and help the educational organizations to get insight into students' educational life. Analysing down such data, on the other hand, can be challenging therefore a qualitative research and significant data mining process needs to be done. Sentiment classification can be done using NLP (Natural Language Processing). For a social network that provides micro blogging services such as twitter, the incoming tweets can be classified into News, Opinions, Events, Deals and private Messages based on authors information available in the tweets. This approach is similar to Tweetstand, which classifies the tweets into news and non-news. Even for e-commerce applications virtual customer environments can be created using social networking sites. Since the data is ever growing, using data mining techniques can get difficult, hence we can use data analysis tools.

Keywords **— Web-text analysis, Data mining, Social network analysis, Human Computer Interaction (HCI), Sentiment classification.**

I. INTRODUCTION

Human behaviour and social phenomena can be more clearly understood as the data on the social media is provided by the users themselves[6].Social media like twitter, Facebook and forums, provides great platform for students to share their daily life experiences in a very informal and a casual manner. This digital information gives a new perspective for educational researchers to understand students feelings regarding education system [1].By analysing these information educational organizations can get insight into students' issues and enhance student employment and achievements. Manual algorithm for analysing this data is not appropriate as it won't be able to handle large amount of data and automated algorithm won't give us in depth meaning of the data. Users may become overwhelmed by collecting the data from the social media platforms. Classification of short text messages is done to avoid this. Traditional classification method such as "Bag-Of-Words" is not efficient as it does not provide sufficient word occurrences. Hence a greedy strategy is used to select the feature set and accordingly to classify it. This new technique effectively classifies the text according to the set of generic classes such as Events, Opinions, Deals and Private messages [5].Sentiment is a view or an opinion expressed by a person but not proved. The study of sentiment is called sentiment analysis which requires the use of NLP to extract information. Using this sentiment analysis technique, a user's mood can be detected. In [3] this sentiment analysis technique can help a user in getting information such as article, news etc. on social media related to a particular sentiment word or tag. Other information extraction tasks such as review summarization can be enhanced using this sentimental analysis value.

In[4] the experiences shared by the users on various topics in twitter can be both negative as well as positive. So some automated algorithm is needed to classify such posts so that the feedbacks are collected without human efforts. The comparative results can be generated with distant supervision. In twitter there are many tweets which many times have emoticons in them and using the twitter API we can easily extract such tweets which is an improvement over the hand label training data which may or may not have emoticons in them. Social media also enable the creation of virtual environment where interested communities form a

specific firm [2]. It also gives new opportunities to such firms in improving their internal operations and collaboration with customers in a more effective way. In analysing social media data researchers may come up with faulty assumptions if qualitative look at the data is not done. Also this data is ever growing and manual analysis is difficult. Thus we require a social media data analysis tool. Using this tool qualitative and large scale analysis can be done [6].

II. LITERATURE SURVEY

In[1] social media provides the open environment where students can share their educational experiences, mindset and problems related to their learning method. Tweets related to engineering students are taken to recognize the issues and problems that are faced by them during their learning practice. Based on this information problems faced by the students can be understood more clearly and the decision makers can come up with new knowledgeable conclusions and help students in solving their problems.

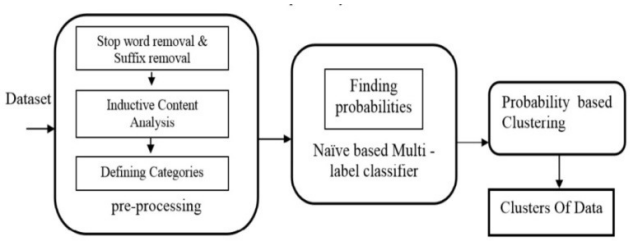

Fig. 1 System Architecture

First the sample of such tweet is taken and then qualitative testing is done on that pattern which is linked to engineering students' educational life. This testing came up with many problems that engineering students were facing such as lack of sleep, Heavy work load, no social gathering. Then using Naïve Bayes multi-label classification algorithm these tweets were classified.

The Naïve Bayes algorithm:

In document di in the trying set, there are Y words Wdi = wi1; wi2;::: ; wiY, and Wdi which is a subset of D. This is done to classify this document into class c or not c. There is independence between all word in this document, and any word wik conditioned on k or k' follows multinomial distribution. Hence, according to Bayes Theorem, the probability that di fit in to category k is

$$p(k|d_i) = \frac{p(d_i|k)\cdot p(k)}{p(d_i)} \alpha \prod_{y=1}^{Y} p(w_{iy}|k)\, p(k),$$

and the probability that di fit into group other than c is:

$$p(k'|d_i) = \frac{p(d_i|k')\cdot p(k')}{p(d_i)} \alpha \prod_{y=1}^{Y} p(w_{iy}|k)\, p(k')$$

Because P (K $|d_i$) p (k'+d_i) ,it normalize the latter two items which are comparative to p(k $|d_i$) and p (k' $|d_i$)to get the actual values of p(k $|d_i$). If p (k $|d_i$) is larger than the probability threshold T, then di fit into category k, otherwise, di does fit into category k. Then this process is repeated for every category.

Probabilistically based clustering algorithms and Kullback-leibler divergence is used. Social media has many methodological complexities like internet slangs, location and the time of post cannot be predicted and handle large amount of data. Manually analysing this large amount of data is impossible and also the use of automated algorithms cannot give us the in depth meaning of the data. Using twitter developers get low latency access to data and it also has a suitable format for analysis. Further analysis of students' data such as images, videos etc. can also be done instead of limiting the analysis to texts only. It can also consider various other majors (not only engineering) and other institutions.

In [2], Large U.S companies use social media platforms like Twitter, Facebook, forums and blogs for creating virtual customer environments (VCEs). These platforms are used to deliver familiar e-commerce applications. Two essential characteristics are defined for effective implementation of VCEs: 1) The firm which attracts the mass of participants from a community and who engage with the other community members 2) The firm which develop processes to benefit from the content created by its customers. Organizations need to create and maintain an infrastructure that supports a community in their virtual customer environments. The purpose of this

infrastructure is to attract and retain a critical mass of participants. Fortune 500's is used to interact with the customers on social media platforms. For gaining full business value from social media, Implementation strategies should be developed by firms based on three elements: mindful adoption, absorptive capacity and community building. Case studies of three Fortune 100 Corporations are used to examine the management of the networks. In [2], the advantage of using social media for VCE is to gain maximum business values. Since it mainly focuses on marketing and uses Social media platforms such as Twitter, Facebook, forums and blogs for the communication. But simply creating an online presence does not ensure that a firm will gain the maximum business value from social media.

Tweets are nothing but small sentences. These tweets contain a large set of user defined hash tags, some of which are sentiment tags that delegate one or more sentiment values to a specific tweet. In [3] a framework is presented which analyses the twitter data and thus sentiment classification is done.

Four different types of feature are used, which are: punctuation, n-grams, words and patterns. Each word which was appearing in a sentence served as a binary feature with weight being equal to the inverted count of this word in the Twitter corpus. Each consecutive word sequence containing 2–5 words was also taken as a binary n-gram feature using the same weighting strategy. n-grams or words which are appearing in less than 0.5% of the training set sentences did not constitute as a feature[3].

For extraction of patterns, words were classified into content words (CWs) and high-frequency words (HFWs). A word whose corpus frequency is more (less) than FH (FC) was considered to be a HFW (CW)[3].Word frequency was evaluated from the training set. A pattern features value was estimated according to one of the following :

$$\begin{cases} \frac{1}{count(p)} : & \text{Exact match – all the pattern components appear in the sentence in correct order without any additional words.} \\ \frac{\alpha}{count(p)} : & \text{Sparse match – same as exact match but additional non-matching words can be inserted between pattern components.} \\ \frac{\gamma * n}{N * count(p)} : & \text{Incomplete match – only } n > 1 \text{ of } N \text{ pattern components appear in the sentence, while some non-matching words can be inserted in-between. At least one of the appearing components should be a HFW.} \\ 0 : & \text{No match – nothing or only a single pattern component appears in the sentence.} \end{cases}$$

$0 \leq \alpha \leq 1$ and $0 \leq \gamma \leq 1$ are the parameters that were used to assign reduced scores for imperfect matches. $\alpha = \gamma = 0.1$ was used in all experiments [3].

In addition to pattern-based features the following generic features were used: (1) Number of "!" characters in the sentence, (2) Number of quotes in the sentence, (3) Number of capitalized/all capitals words in the sentence, (4) Sentence length in words, and (5) Number of "?" characters in the sentence . Normalization was done on these features by dividing them by the (maximal observed value times the averaged maximal value of the other feature groups), hence the maximal weight of each of these features is equal to the averaged weight of a single pattern/ n-gram feature/word.

These all are utilized for sentiment classification and contribution of each type is assessed. As and when new examples are entered into the set, k-nearest neighbours (kNN) classification algorithm is used[3]. Let ti,i = 1.....k be the k vectors with lowest Euclidean distance to v4 with assigned labels as Li,i = 1...k. The mean distance is calculated d(ti,v)for this set of vectors and drop from the set up to five outliers for which the distance was more then twice the mean distance. The label assigned to v is the label of the majority of the remaining vectors. If matching vectors are not found for v, default "no sentiment" label assignment is done.

Algorithm was then tested under several different feature settings: Pn-W-M-Pt+,Pn+W+M-Pt-,Pn+W-M-Pt-, Pn+W+M+Pt- and FULL, where +/– stands for utilization/omission of those feature types:

M:ngrams (M stands for 'multi'),Pn:punctuation, W:Word, Pt:patterns. FULL stands for utilization of all feature types. In this experiment setting, the training set was divided to 10 parts and then a 10-fold cross validation test is executed. Every time, 9 parts is used as the labeled training data for feature selection and also for construction of labelled vectors and the other remaining part used as a test set. The process is then repeated ten times.

The Amazon Mechanical Turk service was used to present the tasks to English-speaking subjects for human evaluation using judges. Each subject was given 50 tasks for Twitter hash tags or 25 questions for smileys. Each set was presented to four subjects. If a human subject did not succeed to provide the required "correct" answer to at least two of the control set questions, he/she was rejected from the calculation. In evaluation, the algorithm is considered to be correct if one of the tags selected by a human judge was also selected by the algorithm. Table 1 shows results for human judgment classification. The agreement score for this task was $\kappa = 0.41$ (agreement was considered when at least one of two selected items are shared). Table shows that the most of the tags selected by humans matched those selected by the algorithm [3].

Table 1
Results of human evaluation. The second column indicates percentage of sentences where judges find no appropriate tags from the list. The third column shows performance on the control set.

Setup	% Correct	% No sentiment	Control
Smileys	84%	6%	92%
Hashtags	77%	10%	90%

Even though the hash tag labels are distinct to data on twitter, the feature vectors that are obtained are not very twitter specific.

In [4], a method was studied which represents a sentiment into positive and negative. This method also helps in cutting the manual efforts that might be used for the same work. The objective is to use tweets and emoticons for distant supervision learning. This method gives high accuracy and works better when combined with machine learning algorithms. In this study, Naïve-Bayes multi label classifier is used which is a simple model. It works well on text categorization. A multinomial Naive Bayes model is used. Class c∗ is assigned to tweet d, where

$$c* = argmac_c P_{NB}(c|d)$$

$$P_{NB}(c|d) := \frac{(P(c)\sum_{i=1}^{m} P(f|c)^{n_i(d)})}{P(d)}$$

In the formula, f represents a feature and ni (d) constitutes the count of feature fi which can be found in tweet d. There are a total of m features. Parameters P(c) and P (flc) can be obtained through maximum likelihood estimates, and then add -1 smoothing is utilized for unseen features.

The multi-label classification problem is broken into multiple single-label classification problems. Other classifiers such as Maximum Entropy (MaxEnt) and Support Vector Machine (SVM) are also used.

In Maximum Entropy the most uniform models that satisfy a given constraint are preferred. These models are feature-based models. We can add features like phrases to MaxEnt and bigrams without worrying about features overlapping. Representation of the model is:

$$P_{ME}(c|d,\lambda) = \frac{\exp[\Sigma_i \lambda_i f_i(c,d)]}{\Sigma_{c'} \exp[\Sigma_i \lambda_i f_i(c,d)]}$$

In this formula, λ is a weight vector, c is the class, and d is the tweet. The weight vectors are used to decide the importance of a feature in classification. If a weight is high it means that the feature is a strong indicator for the class. Theoretically, MaxEnt performs better than Naive Bayes because of its ability to perform feature overlap better. However, in practical sense, Naive Bayes still performs well on range a variety of problems.

In SVM, input data are the two sets of vectors of size m. Each entry in the vector set corresponds to the presence a feature. Let's take an example, with a unigram feature extractor, each of the feature is a single word found in a particular tweet. If the feature is present in tweet, the value is 1, but if absent, value is 0.

Feature extractors are used, such as unigrams, bigrams, parts of speech tags, unigrams and bigrams.When compared to the features of unigram,

naïve Bayes, accuracy improved (81.3% from to 82.7%) and even for MaxEnt (from 80.5 to 82.7). However, SVM had a decline (from 82.2% to 81.6%). It was found that the POS tags were not very useful. The accuracy performance for MaxEnt increased while for Naive Bayes and SVM decreased negligibly when compared to the unigram results.

A framework is build that treats both feature extractors and classifiers as two distinct components. Different combinations of classifiers and feature extractors are tried to give results. Table has the summarization of above results.

Table 2
Classifier Accuracy

Features	Keyword	Naïve Bayes	MaxEnt	SVM
Unigram	65.2	81.3	80.5	82.2
Bigram	N/A	81.6	79.1	78.8
Unigram+Bigram	N/A	82.7	83.0	81.6
Unigram+POS	N/A	79.9	79.9	81.9

Even though these techniques are good enough, accuracy can still be increased by:

1. Classifying the semantics in the tweet by using a semantic role labeller [4].
2. If the domain of tweets is limited, the classifiers may give better accuracy [4].
3. Neutral sentiments in a tweet require proper attention [4].

On Social media, user micro blogs a wide range of topics. Twitter has a limitation of 140 words per tweets so all the tweets on twitter are called as micro blogs [5]. These micro blogs contains very brief information and it is followed by the link for more detailed source of information. Due to this micro blogging users may have large amount of raw data and it becomes very difficult to handle such data. Solution to this problem is classification of short messages. Use of traditional classification methods such as “Bag of words” has various limitations as short messages provide insufficient word occurrences [5]. To overcome this problem a small set of domain-specific features can be used in which texts are classified into predefined set of generic classes. In which the short messages are classified using existing words and then integrated with the meta-information which is available from the other information sources such as Wikipedia and WordNet. Integrating messages with the meta-information helps the automated text classification and hidden topic extraction works more efficiently. Based on the users’ intentions various class labels and feature set are determined. In this another general approach is proposed which is similar to twitterstand that classifies the tweets into news and non-news. The tweets here are classified into categories such as news, events, opinion, deals and private messages based on the authors’ information and features provided in the tweet. By carrying out various experiments it was proved that classification can also be done with high accuracy without using meta-information and the approach proposed here uses the traditional “Bag of words” classification method. Using this approach user will be able to view only limited tweets based on their interest so 8F approach can be used with such set of generic classes. Noise removal techniques must be used as noisier data affects the performance. To achieve higher precision the similarities in search within the classes along with the semantic information that can be collected from the URLs given the tweets can be supported. This will also be useful when twitter is used on the handheld devises was accuracy and performance plays the major role.

In [6], a web based social media data analysis tool called SWAB (Social Web Analysis Buddy) is used. This tool analyses the twitter data which is automatically downloaded using the twitter search API. Various researchers work jointly on the analysis of social media data one after the other. The tool then encapsulates the inputs provided by the researchers. Based on this analysis a classification model is applied for detection and classification on tweets which are of a particular interest. This method reduces manual labour for analysis of large scale data. Three researches

together analysed 3000 engineering student tweets that were in random selected from the tweets that were downloaded using the hash tag #engineeringProblems.

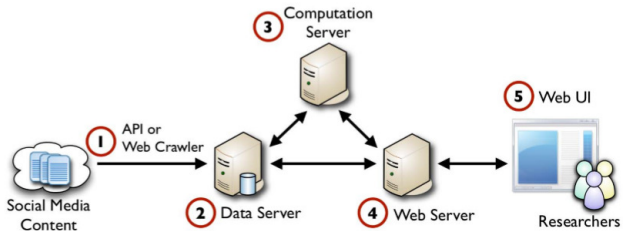

Fig. 2 Architecture of web based social media data analysis too.

The server side of this tools architecture contains the twitter search API, SQL database, naïve Bayes for the computation part and PHP for web-services such as selection of data and sending of results back. The client side contains the interface.

Server Side:

1. Twitter search API is used to collect the twitter data.
2. Through this API data can be queried based on the twitter user, hashtag, posting time, geotag, and keyword. The data which is collected is inserted into a SQL database.
3. The Computation server does many heavy computations such as classification modelling, and inter-rater reliability measures. Naïve Bayes classification is used in SWAB since it is fast, efficient and light and hence very much suitable for web applications.
4. The web server provides the web services which respond to the client side requests. These web services are implemented in PHP and transfer of data is done using JavaScript Object Notation format (JSON).

Client Side:

All the components present at the client side are implemented in JavaScript with the JQuery library, CSS and HTML5. These programs are used in the same machine as that of web server. User interface is designed using: sketch->wireframe->prototype procedures.

The web services and user interface is still not finished and the implementation of this tool is a work in progress.

III. Conclusion:

After referring to the work done by various researchers it can be concluded that analysing data from the social media can give transparent results. Along with the sentiment analysis study and learning of students experiences which will indirectly help educational administration and organizations. After referring the classifier accuracy in [4], we can conclude that Naive Bayes algorithms works best for most of the feature extractors. Also in [1], this algorithm requires less computation time and small amount of pre-defined data. On the other hand in [4], SVM has a limitation of speed and size, and has high algorithmic complexity. In [1] partitioning clustering algorithm has high complexity since even for small number of objects, number of partitions is large.

References

[1] M. Clark, S. Sheppard, C. Atman, L. Fleming, R. Miller, R. Stevens, R. Streveler, and K. Smith, "Academic Pathways Study: Processes and Realities,"Proc. Am. Soc. Eng. Education Ann. Conf. Exposition, 2008.

[2] M.J. Culnan, P.J. McHugh, and J.I. Zubillaga, "How Large US Companies Can Use Twitter and Other Social Media to Gain Business Value," MIS Quarterly Executive, vol. 9, no. 4, pp. 243-259, 2010.

[3] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced Sentiment Learning Using Twitter Hashtags and Smileys," Proc. 23rd Int'l Conf. Computational Linguistics: Posters, pp. 241-249, 2010.

[4] A. Go, R. Bhayani, and L. Huang,"Twitter Sentiment Classification Using Distant Supervision," CS224N Project Report, Stanford pp. 1-12, 2009.

[5] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short Text Classification in Twitter to Improve Information Filtering," Proc. 33rd Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 841-842, 2010.

[6] X. Chen, M. Vorvoreanu, and K. Madhavan,"A Web-Based Tool for Collaborative Social Media Data Analysis,"Proc. Third Int'l Conf. Social Computing and Its Applications, 2013.