Available online at www.sciencedirect.com

ScienceDirect

Materials Today: Proceedings XX (2016) XXX–XXX

www.materialstoday.com/proceedings

PMME 2016

Mining Social Media Data for Understanding Students' Learning Experiences using Memetic algorithm

Swati Patil[a] ,Saroja Kulkarni[b]

[a]*Asst. Professor, VIIT Sr. No-81/82, Kondhwa(Bk) Pune, 411048, India*
[b]*Asst. Professor, VIIT Sr. No-81/82, Kondhwa(Bk) Pune, 411048, India*

Abstract

Now a day's most of students communicate with each other using various social media networks such as Twitter, Facebook, YouTube and what's app. Students shares their opinions, concerns and emotions about the learning process. From these social sites there are large size of unstructured data are generated which consists students important data. To manage this unstructured data are too difficult task, so we use various techniques to solve this problem. In this paper we collect all Engineering students communication from twitter to analyse various problems like heavy study load, negative emotions, lack of social engagement and sleepy problems. Students' comments from twitter are classified into various above problem using Naïve Bayes algorithm. Also we used various algorithms for processing data like stemming, TF-IDF algorithm and cosine similarity. This paper shows that how students share their opinions through twitter and which comments are in which category. Using Memetic algorithm we got the more accurate results.

Selection and Peer-review under responsibility of International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India.

Keywords: Education;problems; computers and education; social networkin; web text analysis

Selection and Peer-review under responsibility of International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India.

1. Introduction

Data mining enables people to discover unanticipated information on unaided basis. This information they can act better understand. The main purpose of our project is to minimize the student's educational problems and also reduce the comment redundancy. Many studies show that social media users may purpose fully manage their online identity to "look better "than in real life.

1.1. Overview of data mining and social media

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Data mining is the process of digging through data and looking meaningful trends and patterns. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration. Data mining can be viewed as a result of the natural evolution of information Technology. Data mining is Iterative Process.

The basic steps are:

1 Data cleaning (to remove noise and inconsistent data)

2 Data integration (where multiple data sources may be combined)

3. Data selection (where data relevant to the analysis task are retrieved from the database)

4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations)

5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user).

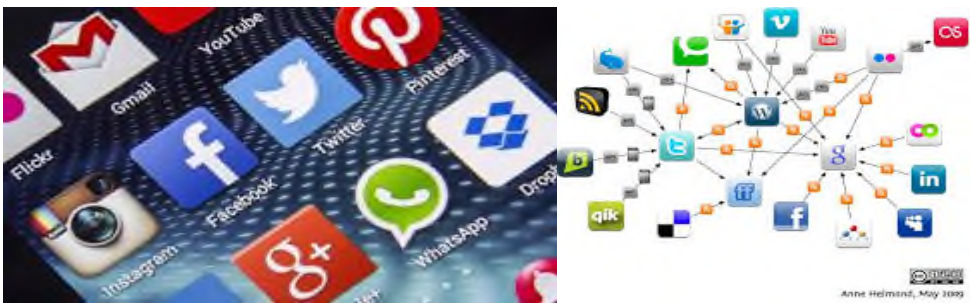

Fig.1. *Social Media*

1.2 Social media mining purpose

Students' informal conversations on social media (e.g. Twitter, Facebook) shed light into their educational experiences opinions, feelings, and concerns about the learning process. Data from such un-instrumented environments can provide valuable knowledge to inform student learning. Analysing 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. In this paper, we developed a workflow to integrate both qualitative analysis and large-scale data mining techniques.

1.3 Goal

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[1].
2. To explore engineering students' informal conversations on Twitter, in order to understand issues and problems students encounter in their learning experiences. There are so many social media sites available in market such as (Facebook, twitter).there are 1 million active user on Facebook and 506 million user on twitter .daily user send 500 tweets and upload 55 million photos. So, it is necessary to mining this large amount of data to avoid complexity. Goal of this study is mining data to understand student's problem because there are 60% of user is none other than students. So based on they tweet we classify their problems in many category. And through this category there are many classification algorithms. Logistic regression is one of them. Regression technique naturally suited to such data.

There are many algorithms available in data mining such as:

1. Naive Bayes
2. Maximum entropy
3. Support vector machine
4. Logistic regression

In this paper we implement Memetic algorithm.

2 Related Work

2.1 Twitter data mining:

2.1.1 Conflation Algorithm

Rules for removing a suffix are given in the form
(condition) S1 → S2 i.e., if a word ends with suffix S1, and the stem before S1 satisfies the condition, then it is replaced with S2.
Example : (m > 1) EMENT → \ Example: enlargement → enlarge

Condition on conflation Algorithm
*S - stem ends with s
*Z - stem ends with z
*T – stem ends with t
v - stem contains a vowel
*d - stem ends with a double consonant
*o - stem ends cvc, where second c is not *w*, *x* or *y* e.g. –wil, -hop
In conditions, Boolean operators are possible e.g. (m>1 and (*S or *T))

2.1.2 Porter stemming algorithm

Conflation Algorithm serves different purposes. Generally, motivation is to achieve an engineering goal rather than linguistic fidelity. This can cause errors in the bag of words mode. Soundex and Porter are very well established and easily available. Porter Stemmers use simple algorithms to determine which affixes to strip in which order and when to apply repair strategies.
Advantage: easy to see understand, easy to implement.

2.1.3 TF IDF Score

Tf–idf, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others. Variations of the tf–idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query. Tf–idf can be successfully used for stop-words filtering in various subject fields including text summarization and classification.

Motivation

Suppose we have a set of English text documents and wish to determine which document is most relevant to the query "the brown cow". A simple way to start out is by eliminating documents that do not contain all three words "the", "brown", and "cow", but this still leaves many documents. To further distinguish them, we might count the number of times each term occurs in each document and sum them all together; the number of times a term occurs in a document is called its term frequency. However, because the term "the" is so common, this will tend to incorrectly emphasize documents which happen to use the word "the" more frequently, without giving enough weight to the more meaningful terms "brown" and "cow". The term "the" is not a good keyword to distinguish relevant and non-relevant documents and terms, unlike the less common words "brown" and "cow". Hence an inverse document frequency factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that we now combine the definitions of term frequency and inverse document frequency, to produce a composite weight for each term in each document.
In other words, assigns to term a weight in document that is
1. Highest when occurs many times within a small number of documents (thus lending high discriminating power to those documents);
2. lower when the term occurs fewer times in a document, or occurs in many documents (thus offering a less pronounced relevance signal);
3. lowest when the term occurs in virtually all documents. At this point, we may view each document as a vector with one component corresponding to each term in the dictionary, together with a weight for each component that is given by. For dictionary terms that do not occur in a document, this weight is zero. This vector form will prove to be crucial to scoring and ranking.

Term frequency

Tf–idf, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection.
Tf–idf is the product of two statistics, term frequency and inverse document frequency.
$Tf\ (t, d) = 0.5 + 0.5\ f_{t}, d) / (\max\{ f_{d}, d: t \in d\})$

Inverse document frequency

The inverse document frequency is a measure of how much information the word provides, that is, whether the term is common or rare across all documents.

Idf (t, D) =log N/(|{dεD:tϵd}|) Where N=total number of documents in the corpus.

3. Collections of Data

In Twitter we use many types of API

- The REST API
  The most common way to access Twitter data is through the REST API. The REST API should meet the needs of most Twitter application programmers
- The Streaming API
  The Twitter Streaming API allows you to receive tweets and notifications in real time from Twitter. However, it requires a high-performance, persistent, always-on connection between your server and Twitter

4. Content Analysis

4.1 Categorize of problems

Categorized into 6 problems like Sleepy Problems, Lack of Social Engagement, Negative Emotion & Heavy Study Load [1].Graph 1.1 shows the categories and their ratio.

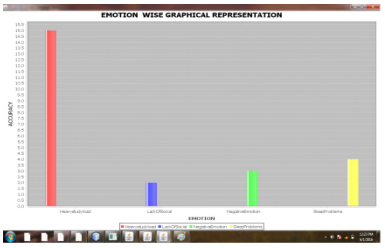

Fig 2. Graph: categories of problems

5. Comparison of Algorithms

5.1. Memetic algorithm

Memetic Algorithm is basically based on Genetic Algorithm. Just we can say , Memetic Algorithm is the combination of Evolutionary Algorithm and individual learning procedure capable of performing local refinements. This means that the algorithm maintain a sets of solutions for the problem. In the context of MAs, the denomination agent seems more appropriate for reasons that will be evident later in this section. When clear from the context, both terms will be used interchangeably. Each individual represents a tentative solution for the problem under consideration. These solutions are subject to processes of competition and mutual cooperation in a way that resembles the behavioural patterns of living beings from a same species.

Procedure Memetic Algorithm
1 **Initialize:** Generate an initial population;
2 **while** stopping conditions are not satisfied **do**
3 *Evaluate* all individuals in the population.
4 *Evolve* a new population using stochastic search operators.
5 *Select* the subset of individuals, Ω_{il} , that should undergo the individual improvement procedure.
6 **for** each individual in Ω_{il} **do**

7 *Perform* individual learning using meme(s) with frequency or probability of f_{il}, for a period of t_{il} .
8 *Proceed* with Lamarckian or Baldwinian learning.
9 **end for**
10 **end while**

5.2 Result analysis

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 [2].

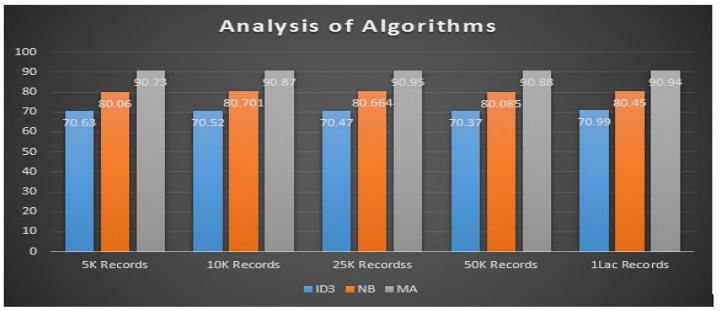

Fig 3:Graph :Comparative Study of ID3 , Navie Bayes and Memetic Classification Algorithms

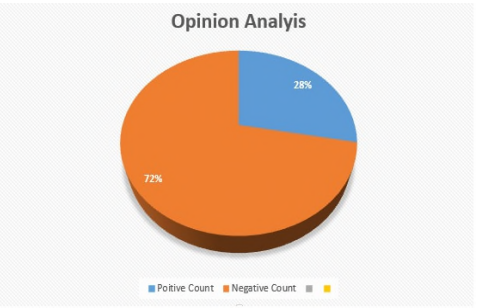

Fig 4: Opinion Generation used in the Memetic Classification.

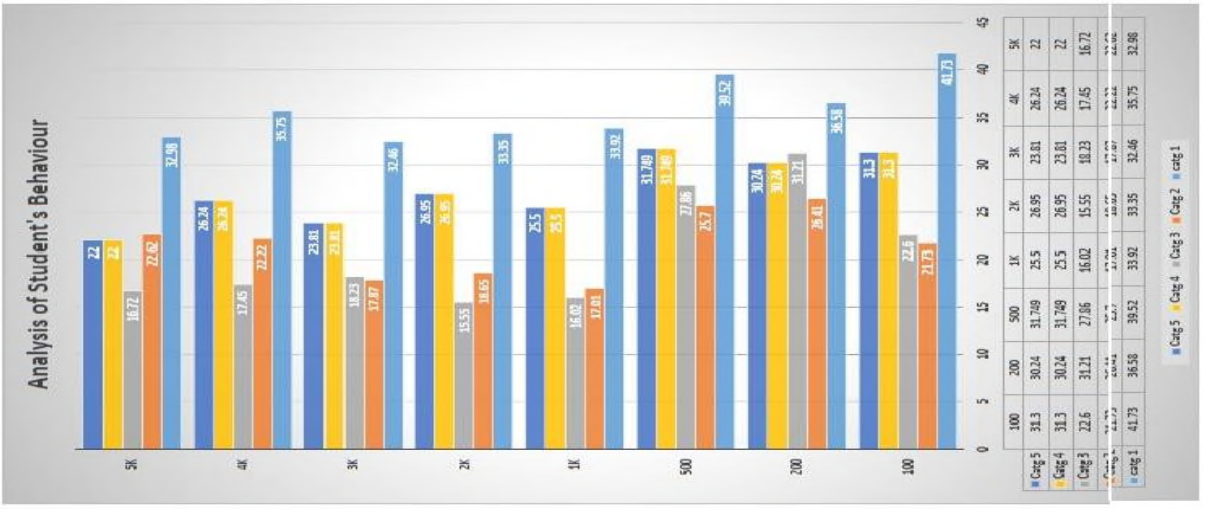

Fig . 5 Classify the student behaviour data for various sizes

6. Future work

Compare Both Good and Bad Things **:** Means compare both to investigate trade-off with which student structure.
*Social Support***:** Future work can be done on why and how student seek and used social media site.
*Detection of Pattern in Digital Media***:** Analyze student generated content other than text.
Counter Terrorism, Privacy concern: Future work can be done to design more sophisticated algorithm for hidden information.

7. Conclusions

Our work is useful for education administrators and policy makers for making them better policy for engineering students which help them for increasing students placements and solving their social problems.

It is initial step to understand students problem without using any traditional system and it may helpful for our society to find at risk students. This attempt helps in informing educational policymakers about student's problems.

8. References

[1] Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan "Mining Social Media Data for Understanding Students' Learning Experiences", 1939-1382 (c) 2013 IEEE.
[2] Natalio Krasnogor, Alberto Aragón and Joaquín Pacheco, MEMETIC ALGORITHMS, School of Computer Science and I.T. University of Nottingham. England.
[3] M. Vorvoreanu and Q. Clark, "Managing identity across social networks," in Poster session at the 2010 ACM Conference on Computer Supported Cooperative Work, 2010.
[4] M. Vorvoreanu, Q. M. Clark, and G. A. Boisvenue, "Online Identity Management Literacy for Engineering and Technology Students," Journal of Online Engineering Education, vol. 3, no. 1,2012.
[5] M. Ito, H. Horst, M. Bittanti, danahboyd, B. Herr-Stephenson,P. G. Lange, S. Baumer, R. Cody, D. Mahendran, K. Martinez,D. Perkel, C. Sims, and L. Tripp, "Living and Learning with New Media: Summary of Findings from the Digital Youth Project,"The John D. and Catherine T. Mac Authur Foundation,Nov. 2008.
[6] D. Gaffney, "#iranElection: Quantifying Online Activism," inWebSci10: Extending the Frontier of Society On-Line, Raleigh, NC,2010.
[7] S. Jamison-Powell, C. Linehan, L. Daley, A. Garbett, and S.Lawson, "'I can't get no sleep': Discussing #insomnia on Twitter,"in Proceedings of the 2012 ACM annual conference on HumanFactors in Computing Systems, 2012, pp. 1501–1510.
[8] 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.