Predict Student Score using Text Mining in English for Librarian Course

Febrianti Widyahastuti[1] and Viany Utami Tjhin[2]
[1]School of Information Technology, Deakin University, 221 Burwood Hwy, Burwood, Victoria 3125, Australia
[2]Department of Information Systems, Bina Nusantara University, Jakarta, Indonesia

Keywords: Text Mining, Prediction, Discussion Forum, Classification, Information Retrieval, Learning Analytics, Learning Management System (LMS)

Abstract: As we know that most information contain text document, studying such issue are promising research areas because many documents need deep learning to discovery new phenomena. This paper aims to identify and discover new knowledge through analysis of text extraction in online discussion forum that capable to predict students' performance by applying text mining from an undergraduate English for Librarians course for one semester in Open University, Indonesia. The result of prediction model in this research can be integrated with the current conventional evaluation process. Additionally, prediction approach can give the best practice that the evaluation method can be predicted using text mining in online discussion forum. In this research, there are two approaches used to predict students' performance: first, incorporating learning material documents and each students' response every week. In this case, algorithm using TF-IDF approach is used to leverage the information from students' response and learning materials about how often words occur in both documents. Second, classifying terms into three categories: students' answer text related to learning material, English meaningful text related to learning material and Indonesian meaningful text related to learning material. The correlation result shows that English meaningful text related to learning material have strong relationship with students' performance.

1 INTRODUCTION

Learning analytics can provide powerful analytical tools from varied sources such as audit logs of students' activities and discussion log interactions in Learning Management System (LMS). The idea has motivated us to focus on useful informational text on online discussion forum logs to find meaningful knowledge using text mining and understanding of students learning progress and behaviour in the learning environment.

The use of text mining in document management become the most promising trends in improving the accuracy and speed of document analysis. As a part of the artificial intelligent form, text mining establishes mapping process of the artificial intelligent at various levels of implementation.

The fact that majority of web data are constructed in unstructured text format that is not automatic and need processes to be understood (Li & Wu 2010); and the intention of many researchers who try to get useful information as well as meaningful knowledge from tremendous amounts of text on online discussion make it necessary to develop innovation of prediction model based on text document on online discussion forum. The result of such kind of research will help the process of acceleration of educational assessment and improving the quality of learning.

The focus of this study was on the prediction of students' performance based on students' response on online discussion forum dataset. The experiment was conducted by involving 69 students enrolled in English for Librarian course. Additionally, the texts used as the basis of analysis were the mixture of Indonesian and English text.

This study aimed to find texts with the highest frequency and whether those texts are related to learning materials. To address such purpose, the data collection and analysis was done by utilizing TF-IDF (Term Frequency and Inverse Document Frequency) method.

Widyahastuti, F. and Tjhin, V.
Predict Student Score using Text Mining in English for Librarian Course.
DOI: 10.5220/0009016900002297
In *Proceedings of the Borneo International Conference on Education and Social Sciences (BICESS 2018)*, pages 83-89
ISBN: 978-989-758-470-1

2 RELATED WORK

Text mining is an extension of data mining—the process of extracting meaningful information from unstructured text (Feldman & Dagan 1995; Hung 2012). Furthermore, (Fayyad, Piatetsky-Shapiro & Smyth 1996) proposed the main purpose of text mining is to find words that can represent the contents of the document so the meaning between documents as well as their relation can be analysed.

Text Mining can be applied in many areas of research such as customer service, E-learning, social networking, bio informatics, trend analysis, security, intelligence, web and email. According to (Hashimi, Hafez & Mathkour 2015). There are five advantages from text mining: helping the extraction of useful information from bulk of data in short time and in efficient way, assisting future aspects prediction based on provided observations and statistics, helping the creation and building the patterns from the provided data which tells us about increasing or decreasing trends, e.g. in business and economy. In addition, Text mining software's also helps the security agencies by monitoring and analysing textual data gathered from internet sources blogs, etc. They can also be used in biomedical databases, where these techniques improve the search of literature.

In text mining process, the pattern of extraction process manifest in the form of useful information and knowledge from a large number of text data sources, such as word documents (.doc files), PDFs, text citations (.txt), and so on.

3 PROPOSED PREDICTION MODEL USING TEXT MINING

The proposed prediction model using text mining is illustrated in Figure 1. As it is seen, there are five stages of process: data collection, pre-processing, feature selection, filtering data (centroid base classifier) and prediction result.

3.1 Data Collection

There are 69 students' enrolled in the English for Librarian course and around 20 students are actively response. The texts are written in mixture between Bahasa Indonesia and English. The whole process of data processing based on text mining approach aimed to get the valuables pattern of prediction obtained from students' response on online discussion.

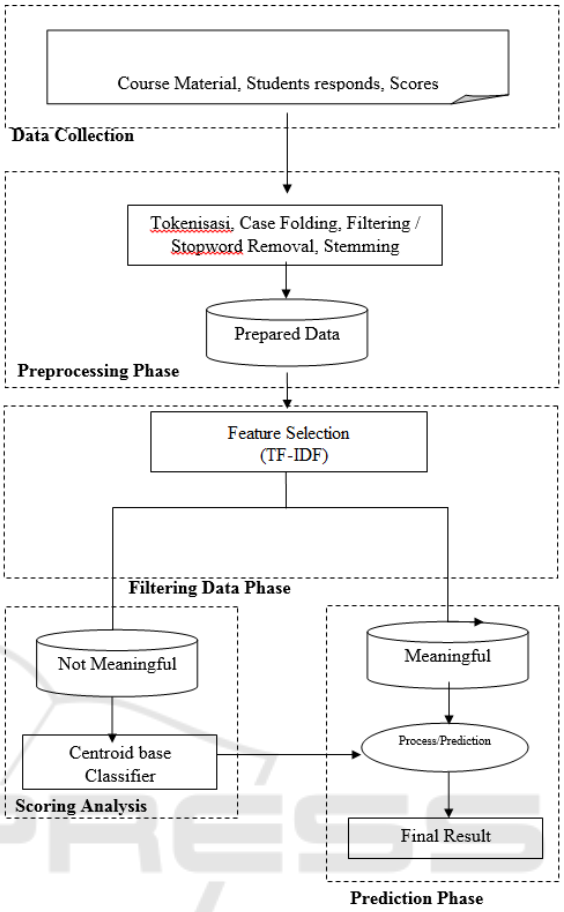

Figure 1: Prediction model using text mining.

3.1.1 Course Material and Student

The text data in English for Librarian course are taken for 8 weeks period of discussion on online forum. In each week, the teachers give the learning material to be discussed. Response from students are the important subject figuring their expression. The Table 1 below displays the captured dataset of learning materials.

Table 1: Capture Dataset Learning Materials.

Meeting Course	Learning Materials
Week 1	Hi students... Here, I will give you the first material to be discussed. Sebagai materi inisiasi 1, saya akan memberikan ulasan mengenai kalimat transitif dan intransitif yang mengacu ke Kegiatan Belajar 2 Modul 1 buku materi pokok Bahasa Inggris untuk Pustakawan. Pertama, yang dimaksud dengan Kata kerja transitif bisa didefinisikan sebagai kata kerja

	yang harus diikuti objek untuk melengkapi maknanya. Mayoritas kata kerja dalam bahasa Inggris termasuk kata kerja transitif.
Week 2	"Hello guys... how are you..?? Hope you are all just fine. This time we'll talk about Adjectives. Adjectives atau kata sifat adalah suatu kata yang diikuti oleh kata benda yang digunakan untuk menambahkan makna terhadap kata benda tersebut.(Wren & Martin, S.Chand & Company, 1974). Adjectives dalam kalimat dapat menjelaskan kata benda (nouns), kata ganti(pronoun) dan linking verbs. Berikut ini adalah contoh adjectives yang berfungsi menjelaskan kata benda: • I have a new car; new adalah adjective yang menjelaskan kata benda "car" • She likes the blue skirt; blue adalah adjective yang menerangkan kata benda "skirt" .....
Week 3	"Hi guys... How time flies... now we're on the 3rd week of this online tutorial. I hope all of you still keep tuning in the discussion here.. Berikut saya berikan materi inisiasi 3 yang mengulas tentang kalimat majemuk dan kalimat kompleks. 1. Kalimat majemuk (compound sentence) Kalimat majemuk adalah kalimat yang terdiri atas dua atau lebih kalimat sederhana atau kalimat yang berisi dua atau lebih independent clause yang dihubungkan dengan kata penghubung (Team of Five: 2001). Untuk menghubungkan dua independent clause tersebut menjadi kalimat majemuk, maka dapat menggunakan kata penghubung (coordinating conjuntion) (Susan Jaderstorm: 2003). .....
Week 8	"Guys... this is the last week of our activity. Please read the following material, and we'll discuss it together, OK... According to the RAT, this time we'll discuss about Gerund. Dalam suatu kalimat atau percakapan, seringkali kita menggunakan bentuk gerund dan infinitiveuntuk mengungkapkan suatu maksud. Gerund adalah bentuk kata kerja yang berakhiran dengan ing seperti blowing, opening, dan having, sedangkaninfinitive adalah bentuk kata kerja yang mengikuti to seperti to blow, to open, dan to have.Gerund disebut juga "a verb noun" karena gerund merupakan kata benda yang berasal dari kata kerja. Perhatikan contoh-contoh berikut ini. It's marvelous to havea 17th birthday party for a girl. Blowingout 70 candles is too hard. We get to open our presents. Opening the presents is so much fun.

Table 2 below displays the students' response in Week 1.

Table 2: Capture Students' Response in Week 1.

Student Name	Students Response
ASEP	Kalimat transitif Arif borrowed a history book from library Kalimat intransitif Mega moves slowly
RIFATUL	"contoh kata kerja transitif;Any wathing a television contoh kata kerja intrasitif:Rifa birthday party tomorrow night" .....

Table 3 show the sample data of students' score. They are taken as the basis of the current result and prediction result comparison. Additionally, In the future, this model can also be integrated with current conventional evaluation process.

Table 3: Students Score.

Student Name	Score		
ASTRI	90	85	87.50
ANDI	85	85	85.00

3.2 Pre-processing Phase

Pre-processing, is the process to prepare the data before the data mining process. Initial data that need to process into text mining are generally still in the form of data that is not fully ready to be processed. The process of data must be done into text format that can easily interpreted and better accuracy of data processing results. All data in the form of excel need to be changed into text. The diagram below illustrates the pre-processing process. Figure 3 below illustrates the pre-processing process.

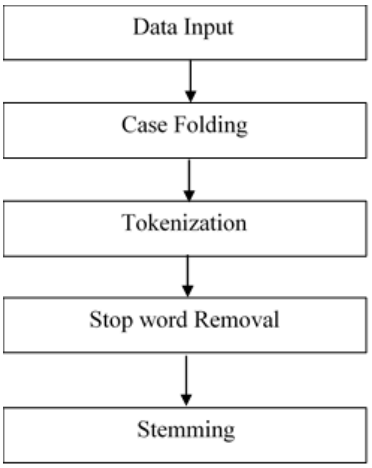

Figure 2: Pre-processing Text Mining Diagram.

The first step in pre-processing is case folding process. This process to uniform all the letters into lowercase, uppercase and eliminate useless characters such as dots, commas, question marks, and so forth. Then the process continued by doing tokenization is to break the sentences, sentences that exist into words or its own words. Next is stop word removal process that is throwing out unneeded words like the word: this, with, the and so on. The words "hi", "here", "i", "you", "the", "to", "as", "1" are removed by the stop word removal process. The example words before and after case folding, tokenization and stop word removal, as shown in Table 4.

Stemming work by removing the end of the word into single term. This may be done by removal of the various suffixes -ED, -ING, -ION, -IONS to leave the single stem DISCUSS. The words were stemmed using Porter's suffix-stripping algorithm. (M.F. 1980).

Table 4: Case folding, Tokenization and stop words.

hi	students	here
i	will	give
you	the	first
material	to	be
discussed	sebagai	materi
inisiasi	1	saya

...	students	....
...	will	give
....	....	first
material	....	be
discussed	...	materi
inisiasi	....	….

3.3 Filtering Phase

The filtering stage are divided into meaningful and no meaningful text. Meaningful text data is data that having correlation with learning materials. Meanwhile, no meaningful text are selected whether text has connectivity with English materials.

In this stage, the data of students' score data in the class are included in the process of prediction analysis of those score. The score of prediction result becomes one of the important factors that text mining with meaningful data can perform prediction..

3.3.1 TF-IDF

The TF-IDF method is a method for calculating the weight of each of the most commonly used words in the information retrieval process. This method is very efficient, easy and has accurate results. It will calculate the value of Term Frequency (TF) and Inverse Document Frequency (IDF) on each token (word / text) in each document in the corpus. At this stage of analysis, the frequency of word with English terms has becomes important comparative factor to predict the value of those words in text mining process. The calculation of the weight of each token in document is done by this formula:

$$Wdt = tfdt * IDFt \quad (1)$$

where :

- d : d document
- t : keyword to t-of keyword
- W : the weight of the d th document against to the t word
- tf : the number of words searched for in a document
- IDF : Inversed Document Frequency

IDF value obtained from IDF: log2 (D/df) where:

- D : total documents
- df : many documents containing the word searched

After the weight (W) of each document is known, then, the process of sorting the value of W based on its level is performed—the greater the level, the bigger degree of similarity it will have towards the keywords and vice versa.

Table 5: Data Input Text Processing with TF-IDF.

Text (Document)	Document Type
D1	Document 1 is the total text document from learning materials used in English learning (as shown in table 1)
D2	Document 2 is the total text students' responds of learning materials using English terms (as shown in table 2)

After pre-processing stage, the data input of text processing in document D1 (document 1) and D2 (document 2) are classified manually as it is seen in table 5. D1 is the total document (text) of the learning materials using English terms. Meanwhile, D2 is a document (text) obtained from students' reply about learning materials.

Table 5 shows the matrix of D1 and D2 of one of student in English Librarian named 'Ade'—her response toward learning material using English terms in week1. Each row of the data processing result represents a word, and the matrix D (i, j) corresponds to the number of occurrences of the word j in document i. The matrix form between terms and documents is shown in Table 6.

Table 6: Example of matrix for Ade (Week 1).

No	Term	D1	D2
1	I	16	3
2	will	6	3
3	give	3	2
4	you	14	3
5	the	32	0
6	first	1	0

Based on the formula of TF-IDF, Table 7 shows the value of DF and IDF for 'Ade'.

Table 7: DF and IDF results.

No	Term	D1	D2	DF	IDF
1	I	16	3	2	0
2	will	6	3	2	0
3	give	3	2	2	0
4	you	14	3	2	0
5	the	32	0	1	0,301029995664
6	first	1	0	1	0,301029995664

The DF shows the two documents (D1 and D2) that contain the text. If both documents (D1 and D2) contain the same text, then the value is 2. However, in the case that only one text in either D1 or D2 contained in, then, the DF value is 1. The formula of IDF can be described as follow:

$$IDF = Log (N/ DF) \quad (2)$$

The DF value=1, the following is the calculation of IDF:

$$IDF= Log (N/ DF) = Log (2/1) = 0,301029995664 \quad (3)$$

After obtaining IDF, The weighted TF-IDF results, then, are shown in Table 8 using above formula.

Table 8: TF-IDF results.

No	Term	D1	D2	Weight D1 (WD1)	Weigh t D2 (WD2)
1	I	16	3	0	0
2	will	6	3	0	0
3	give	3	2	0	0
4	you	14	3	0	0
5	the	32	0	9,63295986125	0
6	first	1	0	0,301029995664	0

3.4 Scoring Analysis

The scoring analysis is done to decide the categorization of meaningful and no meaningful text. The analysis is employed using centroid base classifier algorithm based on TF-IDF result. The results from TF-IDF can be used to process the categorization of data and also sort the amount of text collected by each student. Such step allows us to classify the learning materials into meaningful and no meaningful terms by considering the frequency of the same words used by students in learning materials. Additionally, the meaningful results are broken down into two categories: for English text and Indonesia text.

Table 9: Score analysis by meaningful text.

No	Total	Total Text	Meaningful Text	Score
1	81	23	19	90
2	51	18	13	90
3	48	13	13	90
4	17	6	4	90

3.5 Prediction Results

In this prediction stage, there are three text categories of patterns used as the basis: total text of each students' answer related to learning material, English

meaningful text and Indonesian meaningful text related to the learning material. Table 10 show the results of total text of students' answer in learning materials, English text and Indonesian text. From table 10 displayed, it is seen that there is a sequence of significant English text patterns closely related to the learning materials and score.

Table 10: Result total text, English text and Indonesia text.

No	Total Text	English Text	Indonesian Text	Score
5	276	103	83	89
6	183.	64.	67.	87.5
7	90.	37.	12.	87.5
11	75.	34.	21	87.5
8	66.	34.	10.	87.5
12	55	33.	5.	87.5
26	81	26	23.	85
28	42	26	2	85
27	83	22	24	85
21	29	16	1	85
25	82	15	30	85
24	52	15	9	85
23	57	14	19	85
18	24	11	3	85
17	21	11	2	85
46	16	8	1	0
45	6	5	0	0
42	6	4	1	0
44	5	4	1	0
43	4	3	0	0

3.5.1 Analysis Approach Meaningful English Text

Grouping of the above processed data will provide a rediction scoring model as shown in Table 11.

Table 11: Prediction Score Model.

No	Total Text	English Text	Indonesian Text	Score
Score of 89 is obtained on the Number of Text English is significant from the result of student's answer in the amount of 103				
5.	276	103	83	89
Score of 87.5 is obtained on a significant number of Text English from the student's answer in the 33 -64 range.				
6.	183.	64.	67.	87.5
7.	90.	37.	12.	87.5
11.	75.	34.	21	87.5
8.	66.	34.	10.	87.5
12.	55	33.	5.	87.5
Score of 85 is obtained on a significant number of Text English results from student answers in the range of 11 -26				
26.	81	26	23.	85
28.	42	26	2.	85
27.	83	22.	24.	85
21.	29.	16.	1.	85
25.	82	15	30.	85
24.	52.	15.	9.	85
23.	57.	14.	19.	85
18.	24.	11	3.	85
17.	21.	11.	2.	85
Score of 0 is obtained in a significant number of Text English from the student's answer in range 3 -8				
46.	16.	8.	1.	0
45.	6.	5.	0.	0
42.	6.	4.	1.	0
44.	5.	4.	1.	0
43.	4.	3.	0.	0
No." Indicates the students number				

In general, the patterns of score obtained from text mining by conducting an analysis of the number of meaningful English text which is shown as follows:

1) The score of 89 is obtained from the number of English text for student answer result in amount 103
2) The score of 87.5 is obtained from the number of English text for student answer result in the range 33 -64
3) The score of 85 is obtained from the number of English text for student answer result in range 11 -26
4) A score of 0 is obtained from the number of English Text for the student's answer result in range 3 -8.

If the data equipment is done for the range that is still empty, then the value pattern will be obtained as follows:

1) Number of English text between 103 and above: score > 89 to 100
2) Number of English text at 103: score 89
3) Number of English text in the range 65 - 103: score 87.5 to 89
4) Number of English text in the range 33 -64: score 87.5
5) Number of English text in the range 27 -33: score 87.5 to 85
6) Number of English text in the range 9 -26: score 85 to 0
7) Number of English text in the range less than 3 - 8: score 0
8) Number of English text is less than 2: score not yet defined

Thus, if there are new data that will be processed into the entry, they can easily be predicted (the values) which are obtained by looking at the range that appears and obtained in the data processing above. Example prediction score: Suppose that the

data have 200 points of the amount of English meaningful text, then, the value is defined by pattern 1 or in other words the value obtained is above 89 to 100.

4 CONCLUSIONS

The amount of text that is generated every day is increasing dramatically. This tremendous volume of mostly unstructured text cannot be simply processed and perceived by computers. Therefore, efficient and effective techniques and algorithms are required to discover the useful patterns of information desired. In this case, text mining can be a proper solution to address the issue due to its nature—the task of extracting meaningful information from text, which has gained significant attentions in recent years. Text mining towards learning material and students' performance have correlation with English text meaningful with prediction score model as shown in Table 12.

Table 12: Prediction Score.

No	Number of English Text Meaningful	Prediction Score
1	103 above	Above 89 – 100
2	103	89
3	65 – 103	87.5 – 89
4	33 – 64	87.5
5	27 – 33	85 – 87.5
6	9 – 26	0 – 85
7	3 – 8	0
8	Less than 2	Not yet defined

From the data analysis elucidated before, it is clearly seen that Text mining utilization towards learning material and students' performance have successfully derived the correlation between English meaningful text and students' score by using score prediction model displayed before

The use of Text Mining towards the data of English Language Teaching material in have also been proven effective in predicting students' score. It also gives more objective results compared to subjective conventional assessmet ussually done by education practitioners.

Additionally, the use of Text Mining on discussion forum is also proved to be able to provide prediction model for students as parts of students' assessment and evaluation.

REFERENCES

Fayyad, U, Piatetsky-Shapiro, G & Smyth, P 1996, 'From data mining to knowledge discovery in databases', AI magazine, vol. 17, no. 3, p. 37.

Feldman, R & Dagan, I 1995, 'Knowledge discovery in Textual Databases (KDT)', paper presented to Proceedings of the First International Conference on Knowledge Discovery and Data Mining, Montréal, Québec, Canada.

Hashimi, H, Hafez, A & Mathkour, H 2015, 'Selection criteria for text mining approaches', Computers in Human Behavior, vol. 51, pp. 729-33.

Hung, Jl 2012, 'Trends of e- learning research from 2000 to 2008: Use of text mining and bibliometrics', British Journal of Educational Technology, vol. 43, no. 1, pp. 5-16.

Li, N & Wu, DD 2010, 'Using text mining and sentiment analysis for online forums hotspot detection and forecast', Decision support systems, vol. 48, no. 2, pp. 354-68.

M.F., P 1980, 'An algorithm for suffix stripping', Program, vol. 14, no. 3, pp. 130-7.