Improving The Accuracy of Naïve Bayes Algorithm for Hoax Classification Using Particle Swarm Optimization

[1]Akhmad Pandhu Wijaya, [2]Heru Agus Santoso
Computer science
Dian Nuswantoro University
Email : wijayapandhu39@gmail.com, heru.agus.santoso@dsn.dinus.ac.id.

Abstract— **Hoax news circulation is very widespread which occurs in the media information, both print and online media. For some people hoax news can only appear on the online media. But printed medias also often include hoax news in their published news. In the present era, it is very important providing information with relevant and additional facts otherwise it is categorized as hoax. Therefore, hoax classification approach is needed. This paper focuses on improving the accuracy of hoax classification in textual documents contents. Naive Bayes algorithm is used to train dataset with the use of PSO in the algorithm. Experiment is conducted with the trained model over 600 documents. It shows that feature selection with PSO affects the classification results performed using Naïve Bayes. Accuracy increased from 91.17% without using feature selection, to 92.33% when feature selection is carried out using PSO.**

Keywords— text classification, hoax, news, PSO, Naïve Bayes.

I. INTRODUCTION

Rising news circulation of hoax in various media, both online media and printed news media become a serious problem for social life. Some peoples think that hoax news may only appear on online media, but in fact printed media often share hoax news on their news column. However, hoax news with hate speech content mostly appear in social media. The present era is a reflection of the concerns in the history of human life. With technological advances and the ease of access to information, it makes people very readily accept information and trust eventhough the information may be categorized as hoax with hate speech contents. There are many cases happen in our society related to hate speech, as well as events with facts that do not occur. But as if happening and packed with patterns as best as possible, then it aimed at fostering the spirit so that people believe it because of sedition.

If it is viewed in terms of science and technology, today's era is very important to make improvements of hoax classification problem, especially in the classification of text. With the classification of all text contents, text documents can be grouped based on their parameters. We want to do so to facilitate in obtaining a better structured information for reader. However, there are some technical problems in the classification of them, e.g., on the feature selection section. It can be said that all of the features are not appropriate and can be used as a basis to build a classification model. In fact, not all features can be used as a parameter for hoax classification. To select features that match our expectations and to reduce features that do not fit in the feature selection process, it becomes so important. That is why, this paper focuses on the feature selection process. Beside the society must begin to be educated to realize the danger of news hate speech. With the availability of data and appropriate algorithm, this study hopes able to classify news into decent and organized information in a better manner.

This study aims to develop classification pattern of news hate speech, in the hope that it will improve performance of the existing research. Our proposed approach can make optimal synergy in handling the problems caused by the news as it happens nowadays. While in the field of education that is triggering more complex thoughts in the field of assessment of news hoax, so as to make the intellectuals more concerned with the problems that exist in the community. Especially in the development of news hate speech classification and able to realize human civilization for a better society. Improving the ability of hoax classification is carried out by conducting research in the field of text mining empowered with natural language processing based on the urgent need for improving its performance result.

II. RESEARCH METHODOLOGY

Study that attempts to search or extract data pattern from a collectoion of data from various types resources is called mining. While text mining is extracting data pattern in the form of text that aims to define a set of knowledge. Figure 1 depicts text mining stages which mainly consist of four processes.

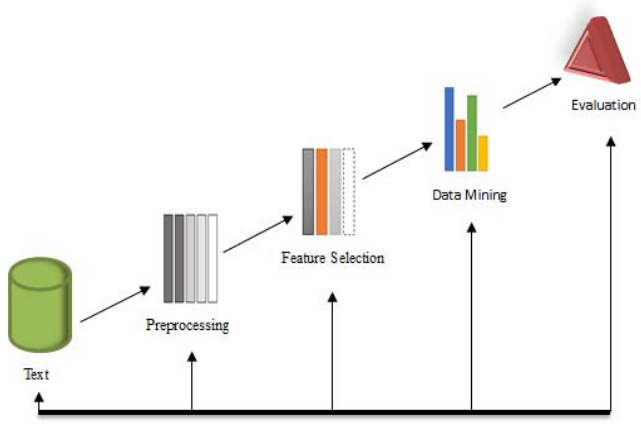

Figure 1 Text Mining Stage

These processes have their respective schemes interconnected with each other because in the results of each process performed will be the main ingredients for the next process. In the text shows that the data that has been collected into a collection of data with the type of text and unknown structures that exist in it. Preprocessing stage will clean up the original document contents which consists of various kinds of characters until the text is ready to be processed. After that, the data becomes more structured as it is necessary in the efforts to clarify the existing documents. On the stage of feature selection, existing features need to be selected in the process of classification into desired class.

2.1 The meaning of text

Indonesian Language has a standard reference in writing that becomes the signs for governance writing. So that it becomes a good communication method. Sentences written on the news as well as information consist of a small section that compiles the word. Basically the word is a unit that is able to stand alone, with one word we can understand what is meant but it is very weak and vulnerable to the question of what the meaning of the word expressed. To express a more complex thought requires the union of words summarized in the sentence, so as to bring up the meaning in accordance with what is meant in the information. To find a sentence and process it with a classification model, clustering and other algorithms, s set of appropriate words is needed that basically has greater impact in the meaning of the sentence or information.

2.2 Hate Speech

The study in this article is centered on hoax news with hate speech content that has a distinctly different pattern of news writing in general. As news hoax often have irregular patterns with inappropriate language, provocative or overstatement and so forth, some of the research that suggested hoax news taken from the media communication either email, news, or chat [9]. The hoax news itself in the dictionary has meanings of jokes, false stories, delinquents, jokes, lies, cheats and so on, and hoaxes have synonyms (practical joke, joke, jest, prank, trick). While in Kamus Besar Bahasa Indonesia (KBBI), hoax is defined as "false news", then hoax can be translated with simple understanding.

2.3 Text Preprocessing

In the text preprocessing stage, cleaning unimportant characters is performed in the pre-processing stage, This stage processes the data until it becomes ready and meaningful. In general, documents have diverse characteristics and have various deimensions, the characteristics of which are meant documents with nois or disturbing parts, and the layout of the text on the documents are unstructured. The structure of the document becomes very important to introduce the pattern so as to characterize a particular class. At this stage there are several steps that must be carried out such as case folding, tokenizing, filtering, and stemming.

a. Case Folding

The consistency of writing documents using capital letters is not always met. Then the capital letter used must be equated with eliminating all the writing with a capital letter by case folding process as depicted in Figure 4. As in the word "TESIS", "Thesis", or "TeSis, when they are input with similar writing model, they will still be read as " thesis ".

b. Tokenizing

The tokenizing process is where text data in the form of paragraphs, sentences, or words are processed with the aim of dividing into specific parts. It can be said they will be converted in to a set of unit of words. Tokezing terminology is a process of replacing the value of the original value into a value called "token" [10]. If the dataset contains a sentence "classification text on hoax news document" it will generate a set of words consits of "classification", "text", "on", "document", "news", "hoax". That way the data that was composed into a paragraph or

sentence will be the units of words that will be the element of document processing. Benefits of tokenizing is simplify the process of classification, with separation of the word elements into the input material becomes more concise. The example of tokenizing process is depicted in Figure 5

c. Filtering

In this section is often also called a stopword list or stop word removal. And this process is a standard operation which is performed on the data in the form of text that aims to take the words result of tokenizing. It then become a representation of documents that often called as a set of features. Technically, the processing using the stop list model allows the user to create a word dictionary as the basis of disposing of unused words on the dataset. Figure 6 and Figure 7 descibe the process of filtering.

ada	apabila	baik	benarkah	berkenaan
adalah	apakah	bakal	benarlah	berlainan
adanya	apalagi	bakalan	berada	berlalu
adapun	apatah	balik	berakhir	berlangsung
agak	artinya	banyak	berakhirlah	berlebihan
agaknya	asal	bapak	berakhirnya	bermacam
agar	asalkan	baru	berapa	bermacam-macam
akan	atas	bawah	berapakah	bermaksud
akankah	atau	beberapa	berapalah	bermula
akhir	ataukah	begini	berapapun	bersama
akhiri	ataupun	beginian	berarti	bersama-sama
akhirnya	awal	beginikah	berawal	bersiap
aku	awalnya	beginilah	berbagai	bersiap-siap
akulah	bagai	begitu	berdatangan	bertanya
amat	bagaikan	begitukah	beri	bertanya-tanya
amatlah	bagaimana	begitulah	berikan	berturut
anda	bagaimanakah	begitupun	berikut	berturut-turut
andalah	bagaimanapun	bekerja	berikutnya	bertutur
antar	bagi	belakang	berjumlah	berujar
antara	bagian	belakangan	berkali-kali	berupa
antaranya	bahkan	belum	berkata	besar
apa	bahwa	belumlah	berkehendak	betul
apaan	bahwasanya	benar	berkeinginan	betulkah

Figure 2 a set of features

d. Weighting

Large-scale documents need to be processed with the appropriate scheme to generate each weight of the object in the form of words, phrases or units in a document that has been processed using indexing. With the process of indexing, it makes the data to have a priority scale based on the numbers of obtained words or features in accordance with their weight that has been determined.

1. Term Frequency (TF)
   The document has a term that will be the determinant of classification. This term is a vector format that can be understood by the computer and the vector frequency is very influential [11]. The frequency of occurrence of the term is declared with Term Frequency (TF). With frequency, set of terms exist in documents can be analyzed with the parameters of the number of terms that appear. So that the value of the document adjustment depends on the larger weights. Term Frequency has several categories that are used as a reference to formulate such a model:
2. Binary TF
   Focused only on the existence of a term or a word on the document, whether the term is contained in the document or not. If the term is on the document, it will be assessed with (1), otherwise it will be given a value (0).
3. Raw TF
   The number of occurrences of a term in a particular document becomes an existing value in TF-IDF. The value will appear if for example there is one term appearing 10 times in the document then the value obtained is 10.
4. Log TF
   Existing documents and slightly contain the desired terms need to be selected so that the frequency of the terms in document decreases.
5. Normalized TF
   It is the ratio of the number of terms / frequency of a term to the frequency of all terms in the document.

The proposed approach is formulated to calculate the value of TF and IDF on every word. The weight calculation on each token in the document is spelled out in the formula :

$$W_{dt} = tf_{dt} * IDF_t \quad (1)$$

5. Inverse Document Frequency (IDF)
   It is the calculations performed to reduce the weight of a term scattered across most documents and make the search process inhibited. The parameter that appears is the less number of documents containing a certain terms will has greater its idf value.

$$IDF_j = \log\left(\frac{D}{df_t}\right) \quad (2)$$

Variables in the formula (1) and formula (2) are explained as follows:

W_{dt} : The word weights d in document t

t_{dt} : The number of occurrences of the word d in document t

D : Number of documents

df_t : Number of documents t containing the word d.

2.4 Feature Selection

Categorization with features in large numbers of collection of documents has its own difficulties in text classification. High feature dimension is very influential in the process of classification because there may exist duplication and redundancy of the feature. The PSO method is an approach that runs on the basis of related interaction patterns and takes into account the quantity factor in the form of population as the main basis of research with large data. This model was initiated by Kennedy and Eberhart in 1995 [2]. The basis of the PSO concept is to represent the problem as a particle present in a collection (swarm), vector is represented as $x_t = (x_{t1}, x_{t2}, \dots, x_{tD})$, where D is a measure that reflects the search space and then the space is entered by the particle that will find the best solution from the search on that variable. The particles in question can be represented as :

$$v_t = (v_{t1}, v_{t2}, \dots, v_{tD}).$$

3. Classification

The process for defining models with different concepts or data classes to provide data estimates refers to the available classes. This process is done after the preprocessing step. The working concept of Naïve Bayes Classifier is with probability on the dataset. Then, machine classifies with this model to determines the probability of the document against a particular class, illustrated that a document D associated with the given class C is $P(C|D)$ [3][4].

$$P(C|D) = \frac{P(D|C)\,P(C)}{P(D)} \tag{3}$$

Where:

P = Probability
C = Class
D = Document

The probability of a document against a given class depends on several variables that become benchmarks, i.e., class probability to document, class probability, and document probability.

4 Evaluation

It is necessary to test and evaluate the proposed model, to know the success rate. The evaluation metrics used in this study are: accuracy, recall, and precision. The value of each metrics are obtained form the so called confusion matrix which is produced during classification process. From the proposed evaluation model, then there will be parameters to know the success, but it is not possible that the results do not match the expectations because the results obtained depends on several factors in it. While the calculation of precission uses the number parameters.

III. DISCUSSION

Feature selection approach design consists of training and testing set. The start-up phase or the training of data is done with preprocessing process which produces the data in the form of features and frequency. Based on the weight generated by preprocessing in the form of weighted values for later selection of features using the PSO, then the resulting feature will be ready to do further process of classification.

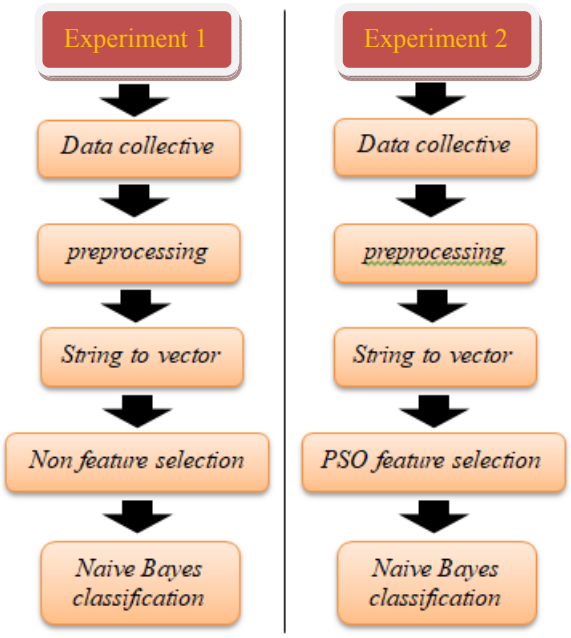

Figure 3 Research Steps

3.1 Data Collection

All of hate speech hoaxes and real data are collected from trusted sources. For the hoax category, the dataset collection is taken from a portal which is specifically acknowledged to present hoax information i.e., https://www.turnbackhoax.id. The data is supplemented with supporting data in the form of text from social media containing hate speech, which is taken from research from Uiversitas Indonesia [13]. The study used data from twitter containing hate speech, as shown below.

NARASI
"Satu lagi leting kita putra terbaik bangsa..alumni SEBA POLRI SINGARAJA (SEBA5) 1986/1987 PLETON I jadi calon PANGLIMA TNI.BRAVO SEBA V SINGARAJA.baru setahun tugas di res sumba ntt, lulus AKABRI AU.".

Figure 4 example of hoax dataset

JAKARTA – Setelah stagnan di Februari, pemerintah mengumumkan perubahan harga bahan bakar minyak (BBM)
Sedangkan harga minyak tanah masih Rp 2.500 per liter, begitu juga dengan solar yang tetap dijual Rp 6
Kepala Pusat Komunikasi Publik Saleh Abdurrahman mengatakan, kenaikan dikarenakan harga minyak di pasa
Jadi, harus ada penyesuaian untuk harga premium. "Demi menjaga kestabilan sosial ekonomi pengelolaan h
Terhitung sejak 1 Maret 2015 pukul 00.00 WIB, untuk premium RON 88 di wilayah luar Jawa, Madura, Bali
Harga sebelumnya Rp 6.600 per liter naik menjadi Rp 6.800 per liter.

Untuk daerah Jamali, Direktur Pemasaran Pertamina Ahmad Bambang menyebut, premium yang sebelumnya Rp 6
Namun, seperti biasa, harga di Pulau Bali biasanya lebih mahal karena pajak yang diterapkan lebih ting
"Kami memahami kondisi masyarakat. Harusnya naik Rp 400 (jadi Rp 7.100, red)," katanya.
Saleh menyebut, harga MoPS Premium mengalami kenaikan menjadi USD 55-70 per barel.
"Kenaikan MoPS sepanjang Februari sebenarnya cukup signifikan. Tapi, Pemerintah tidak menaikkan harga
Untuk solar, harga MoPS sepanjang Februari antara USD 62-74 per barel. Ketidakstabilan harga, lanjut Si
juga terkait pertentangan pelaku pasar minyak dalam menyikapi konflik di Libia. (dim/dio)

Figure 5 example of real dataset

3.2 Preprocessing Dataset

Data processing begins with the process of breaking the data structure into small particles / sentences from the original data with large populations. Then, they are transformed into vectors with different value on each feature, because the frequency that appears different in number. At the beginning of this experiment, an event that appears as a difference in the number of features appears, because this feature is uneven.

Table 1 Example of hoax hate speech

Authentic hoax document
'*Valid* Sangat luar biasa zholim….! Masih ingat musibah Crane di Masjidil Haram..? Kerajaan Saudi bertanggung jawab dan akan menyantuni para korban, tnyata sudah cair sejak lama tapi belum di salurkan kepada korban oleh KEMLU , dengan las an masih verifikasi…! Brp lama verifikasinya pak ..? Sampai kapan itu diverifikasi? Sementara Kafir Harbi , Syiah dan kelompok munafik terus menerus membuat fitnah kepada Arab Saudi. Bantu Share…!
After processed
valid sangat luar biasa zholim masih ingat musibah crane di masjidil haram kerajaan saudi bertanggung jawab dan akan menyantuni para korban tnyata sudah cair sejak lama tapi belum di salurkan kepada korban oleh kemlu dengan alasan masih verifikasi brp lama verifikasinya pak sampai kapan itu diverifikasi sementara kafir harbi syiah dan kelompok munafik terus menerus membuat fitnah kepada arab saudi bantu share

Above table shows that a change of data structure during preprocessing, to get the structure as above. The beginning process are tokenizing and case folding. After the processes are done, then the data must be changed in the form of a matrix to be further processed as presented in Table 2.

Table 2 Matrix of feature weights

Fit/doc	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10	Doc11 ?
abua	0	0	0	6,64	0	0	0	0	0	0	
aburizal	1,73	0	3,47	0	3,47	0	0	0	0	0	
aceh	0	0	0	0	0	0	0	9,96	0	0	
adil	4,64	0	0	0	0	2,32	0	0	0	0	
administrasi	0	0	0	3,32	0	0	0	0	0	0	
adu	0	0	0	0	0	0	0	0	0	3,32	
adzab	0	0	0	0	0	0	0	0	0	3,32	
aneh	0	0	0	0	0	3,32	0	0	0	0	
apalagi	0	0	0	0	0	0	6,64	0	0	0	
firaun	0	0	0	0	0	0	0	0	3,32	0	
Kelas	real	real	real	real	real	hoax	hoax	hoax	hoax	hoax	???

3.3 Results Analysis

a. Beginning with the initialization of the dataset and calculating the entire dataset, the number of hoax and real datasets is explained below.

Number of datasets : 600
Number of hoax data : 300
The number of real data : 300

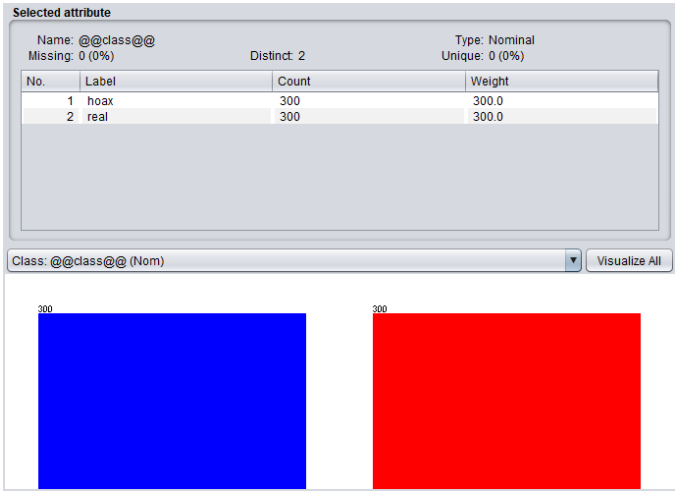

Figure 6 Kind of Class

b. Calculation of weights on the training dataset is collected by documents, which is an equal distribution of each class. Features are selected as a fraction of the existing set of datasets.

c. Naive Bayes experiment with or without PSO are done. Based on two types of experiments above, which are Naive Bayes algorithm without and with feature selection using PSO, the result is promising. The experiment is conducted over 600 documents. It shows that feature selection with PSO improve the classification results performed using Naïve Bayes. The performance increased from 91.17% to 92.33% of accuracy when feature selection is carried out using PSO as depicted in Figure 16.

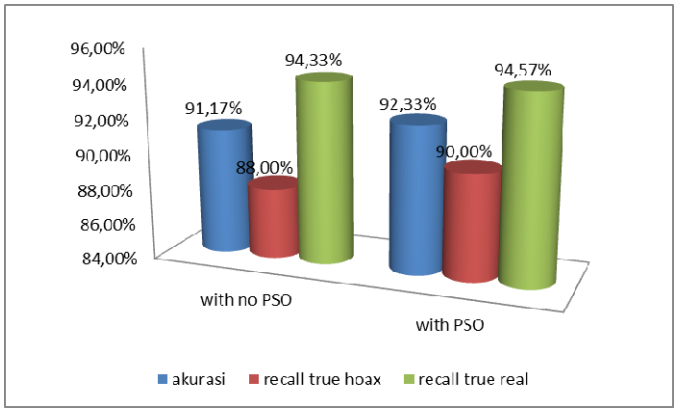

Figure 7 Improved accuracy

IV. RESULT

A. Conclusion

From the discussion as presented in the previous chapter, the authors can draw the following conclusions. Research that has been done and tested, it can be concluded that text feature selection can be done by using Particle Swarm Optimization algorithm and continued by Naïve Bayes classification process. The result indicates that the performance of hoax classification improved by using feature selection.

B. Suggestions

The suggestions that may be given in this study for further development in order to improve the quality and functionality of this document classification method are as follows: Improved text processing and identification and developed the preprocessing stage of the dataset by selecting more text features that are deemed necessary in the dataset to improve the news classification process.

THANK-YOU NOTE

We would like to thank the Government of Indonesia for the funding in the publication of this preliminary study, and also the referees for useful and constructive suggestions.

REFERENCES

[1] J. Vashishtha, "Particle Swarm Optimization based Feature Selection," vol. 146, no. 6, pp. 11–17, 2016.

[2] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1656–1671, 2013.

[3] H. Shimodaira, "Text Classification using Naive Bayes," no. 4, 2015.

[4] S. L. Ting, W. H. Ip, and A. H. C. Tsang, "Is Naïve bayes a good classifier for document classification?," *Int. J. Softw. Eng. its Appl.*, vol. 5, no. 3, pp. 37–46, 2011.

[5] M. Vuković, K. Pripužić, and H. Belani, "An intelligent automatic hoax detection system," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5711 LNAI, no. PART 1, pp. 318–325, 2009.

[6] J. C. Hernandez, C. J. Hernandez, J. M. Sierra, and A. Ribagorda, "A first step towards automatic hoax detection," *Proceedings. 36th Annu. 2002 Int. Carnahan Conf. Secur. Technol.*, pp. 102–114, 2002.

[7] J. Meng, H. Lin, and Y. Yu, "A two-stage feature selection method for text categorization," *Comput. Math. with Appl.*, vol. 62, no. 7, pp. 2793–2800, 2011.

[8] Y. Li, A. Algarni, M. Albathan, Y. Shen, and M. A. Bijaksana, "Relevance feature discovery for text mining," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 6, pp. 1656–1669, 2015.

[9] E. Rasywir and A. Purwarianti, "Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin," *J. Cybermatika*, vol. 3, no. 2, pp. 1–8, 2015.

[10] R. Tokens, "Tokenization Product Security Guidelines – Irreversible and Reversible Tokens," no. April, pp. 1–84, 2015.

[11] M. Liu and J. Yang, "An improvement of TFIDF weighting in text categorization," *Int. Conf. Comput. Technol. Sci.*, vol. 47, no. Iccts, pp. 44–47, 2012.

[12] I. Ahmad, "Feature Selection Using Particle Swarm Optimization," vol. 2015, no. 4, pp. 231–238, 2015.

[13] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate Speech Detection in the Indonesian Language: A Dataset and Preliminary Study," *9th Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS 2017)*, no. October, 2017.